DOCUMENT RESUME

ED 440 107 TM 030 630

AUTHOR Froese-Germain, Bernie
TITLE Standardized Testing: Undermining Equity in Education. Report Prepared for the National Issues in Education Initiative = Les Tests Standardises: Atteinte a L'Equite en Education. Rapport Prepare dans le Cadre de la Campagne des Dossiers Nationaux en Education.
INSTITUTION Canadian Teachers' Federation, Ottawa (Ontario).
ISBN ISBN-0-88989-326-8
PUB DATE 1999-11-00
NOTE 153p.
AVAILABLE FROM Canadian Teachers' Federation, 110 Argyle Avenue, Ottawa, Ontario K2P 1B4, Canada. Tel: 613-232-1505; Fax: 613-232-1886; e-mail: info@ctf-fce.ca. Web site: http://www.ctf-fce.ca.
PUB TYPE Books (010) -- Reports - Descriptive (141) -- Multilingual/Bilingual Materials (171)
LANGUAGE English, French
EDRS PRICE MF01/PC07 Plus Postage.
DESCRIPTORS Academic Achievement; *Accountability; *Achievement Tests; Curriculum; Educational Assessment; Elementary Secondary Education; *Equal Education; Foreign Countries; Performance Based Assessment; *Standardized Tests; Test Bias; *Test Use
IDENTIFIERS Canada

ABSTRACT

Changes in assessment policy have resulted in a proliferation of testing in Canadian schools. Advocates claim that more testing will result in greater accountability, but this book states that there is little evidence to support standardized testing from a pedagogical perspective. It examines the effects of standardized testing on educational equity, emphasizing the impact of testing bias and the misuse of test results. It also discusses performance-based assessment and the politics of testing. Some general observations on the assessment of student learning are presented, and some principles to guide decision making in the areas of assessment, evaluation, and curriculum are suggested. The chapters are: (1) "Introduction"; (2) "Problems with Standardized Testing--An Overview"; (3) "Testing and Inequity in Education"; (4) "A Few Words about Authentic Assessment"; (5) "Politics of Testing"; and (6) "Some General Observations." Appendixes contain a summary of problems with standardized testing, some principles for fair student assessment, and the principles of the Canadian Teachers Federation on curriculum, assessment, and evaluation. (Contains 68 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
G Gillis
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 440 107

STANDARDIZED TESTING

UNDERMINING EQUITY IN EDUCATION

Bernie Froese-Germain

TM 030630

BEST COPY AVAILABLE

2

Standardized Testing:
Undermining Equity in Education

Report prepared for
the National Issues in Education Initiative

by Bernie Froese-Germain
Canadian Teachers' Federation

November 1999

For more information please contact:

Canadian Teachers' Federation
110 Argyle Avenue
Ottawa, ON K2P 1B4
Tel: (613) 232-1505
Fax: (613) 232-1886
E-mail: info@ctf-fce.ca
Internet address: http://www.ctf-fce.ca

The Canadian Teachers' Federation was created in 1920 to ensure that teachers' views were taken into account whenever any authority outside the provincial jurisdiction was considering action which would affect teachers, or impinge on their work with students. Active at the interprovincial, national and international levels, CTF is the national voice of teachers in promoting quality education, the status of teachers and equality of opportunity through education. CTF also coordinates and facilitates the sharing of ideas, knowledge and skills among its provincial/territorial Member organizations which, collectively, represent over 240,000 teachers.

ISBN: 0-88989-326-8

© *Canadian Teachers' Federation, 1999*

The production of this publication including translation, layout and printing was made possible through the collaborative efforts of CTF staff.
Cover Design by Prosar Canada Inc.

Table of Contents

Preface i

Introduction 1
Problems with Standardized Testing – An Overview 3
Inability to assess student learning and development 5
Other problems 12

Testing and Inequity in Education 13
Inappropriate use of tests 13
High-stakes, measurement-driven accountability 18
Bias in testing 23

A Few Words About Authentic Assessment 31

Politics of Testing 35
International comparisons 35
Testing, schools and market competition 38
'Pan-Canadianization' of curriculum and assessment 40
The bigger picture of educational reform 45

Some General Observations 49

Appendix A –
Summary of Problems with Standardized Testing 53
Appendix B –
Principles for Fair Student Assessment Practices for Education in Canada (Joint Advisory Committee) 55
Appendix C –
CTF Principles on Curriculum, Assessment and Evaluation 57
Principles on Curriculum 57
Principles on Assessment and Evaluation 58

References 61

Preface

Canadian schools are awash in student testing. Changes in assessment policy have increased standardized testing at provincial, national and international levels, introduced testing at more (and earlier) grade levels, increased the reporting of test results and attached more significance to those results. Advocates claim that more testing will result in greater school accountability – testing seems to have become synonymous with accountability.

In this context of an increased focus on testing as an instrument of educational reform, CTF's *Standardized Testing: Undermining Equity in Education* argues that there is little evidence to support standardized testing from a pedagogical perspective and looks at the effects of standardized testing on educational equity. Problems with standardized testing are well-documented. They stem from the nature of the tests and the way in which tests and their results are used. This report considers some of these problems, with particular emphasis on the impact of testing bias and the misuse of test results on equity in education. It also discusses performance-based assessment, examines the politics of testing, offers some general observations on the assessment of student learning, and provides principles developed by educators to guide decision-making in the areas of assessment, evaluation and curriculum.

Intended to increase awareness of critical issues related to the impact of standardized testing, this book will be a useful resource for teachers at all levels as well as administrators, researchers, policy-makers and anyone with an interest in student assessment and evaluation.

Bernie Froese-Germain is a researcher with the Canadian Teachers' Federation in Ottawa.

Introduction

Addressing a Canadian Teachers' Federation workshop on curriculum and evaluation in 1998, Benjamin Levin, Dean of Continuing Education at the University of Manitoba, provided an overview of the changes occurring across the country with respect to educational policy.[1] On the curriculum side, he noted that policy changes include an increased focus on the 'basics', fewer options and more restrictive time allocations, the development of regional and Pan-Canadian curricula, and shorter implementation times for new curricula. Changes in assessment policy include more provincial testing, more testing at the national and international levels, increased public reporting of test results despite the generally poor quality of this reporting, and more importance given to the results. Canada is not alone in placing an increased focus on testing. According to Levin (1998),

> an emphasis on standards, accountability and testing has been a feature of reforms in many countries. Almost everywhere we find more large-scale testing of students and more reporting of the results of these tests than was the case a few years ago. Increasing national assessment is complemented by more and more international assessment and in both cases the results are used more overtly for public comparisons. (p. 133)

[1] The National Issues in Education Workshop on "Curriculum and Evaluation: Teacher Organizations and Pan-Canadian Developments" was convened by the Canadian Teachers' Federation in January 1998 to bring the leaders of teacher organizations together to discuss recent developments in curriculum and evaluation, particularly their increasingly "Pan-Canadian" nature. Workshop participants developed statements of principle on curriculum, assessment and evaluation which are contained in Appendix C at the end of this report.

While advocates claim that standardized testing will increase the accountability of teachers and schools, there is little evidence to support such claims. There is, however, a good deal of evidence documenting numerous problems associated with standardized testing (Meaghan & Casas, 1995d, p. 46). Problems stem not just from the nature (form and content) of the tests but perhaps more importantly, from the way in which tests and their results are used. This report considers some of these problems, particularly the impact of testing bias and the misuse of test results on educational equity. It will also briefly discuss authentic or performance assessment, address standardized testing from a political perspective, and offer some general observations on the assessment of student learning and development.[2] Appendices are included which contain principles developed by educators to guide decision-making in the areas of assessment, evaluation and curriculum.

The numerous references in the report to testing in the U.S. reflect the extensive experience Americans have had with this practice, experience which holds valuable lessons for Canada. It is estimated that American public schools administered more than 200 million standardized tests over the 1986-87 school year (Medina & Neill, 1990). According to Barlow and Robertson (1994), "if testing created excellence, American students, the 'most-tested' students in the world, should be winning the education sweepstakes" (p. 117); instead, the U.S. consistently scores low on large-scale international assessments (Earl, 1995, p. 5). Research conducted over the past two decades confirms that standardized tests have had "many negative consequences for the quality of American schooling and for the equitable allocation of school opportunities." (Darling-Hammond, 1991, p. 220)

[2] While the terms *assessment* and *evaluation* are often used interchangeably (as is the case in this report), there is a subtle yet important distinction between them. Assessment is defined as the gathering of reliable information pertaining to students' knowledge and understanding including social, economic and educational factors and resources that influence student and system performance. Evaluation is the process of making judgements, based in part on assessment data.

Problems with Standardized Testing – An Overview[3]

Developed during the First World War to identify army recruits who could be trained as military officers, standardized tests were a means of spreading the candidates out in order to spot those best suited for leadership roles (Casas & Meaghan, 1995, p. 18).[4] Traub (1994) defines standardized achievement tests in the following way:

> Generically, an achievement test is designed to assess the knowledge and understanding a student has acquired of a school subject. A standardized achievement test is further defined by its being given and scored in the same way, whenever and wherever it is used. Standardization means that the scores of all students tested can be fairly compared, one against the other ... the essential requirements are that the conditions of administration and scoring be the same for all the students who are tested so that their scores can be compared. (p. 5)

[3] A summary of some of the major problems with standardized testing is contained in Appendix A.

[4] The U.S. Scholastic Aptitude Test (SAT), developed in the 1920s, was based on the Army Alpha test, a test used to select potential officers from among army recruits (Lemann, 1995). For a brief history of testing in general (including the introduction in the 18th century of ranking and the invention of the quantified mark), see Madaus & Kellaghan (1993).

They are often referred to as norm-referenced tests because an individual's scores are compared to the scores of a larger representative group (the "norm"). The test format is frequently multiple choice, making it easier to mark with the assistance of computers. Aside from advances in electronic scoring, analysis and reporting of results, many standardized achievement tests have changed very little over the past fifty years despite claims that they have become more "scientific" (Medina & Neill, 1990, p. 13) and despite advances in our knowledge about what makes for good teaching and effective learning.

A substantial body of research has documented numerous flaws in the construction, validation, administration and use of standardized tests. According to Medina and Neill of FairTest (the National Center for Fair & Open Testing in Massachusetts),

> [s]tandardized tests are constructed in ways that often guarantee biased results against minorities, females and low-income students. Test results are evaluated and scored in ways that are often at odds with modern theories of intelligence and child development. The test validation process is often inadequate and far from objective. Many tests are administered in an environment that undermines any claims they may have to being "standardized". Even those that adhere to "standard" administration practices may be biasing the results against minorities, low-income students and females by using examiners who are unfamiliar to the test-takers and by using tightly timed tests. These flaws undermine testmakers' claims of objectivity and produce results that are inaccurate, unreliable or biased. Ultimately, many tests fail to effectively measure test-takers' achievement, abilities or skills. (p. 8)

Inability to assess student learning and development

Although standardized tests may be useful for sorting and ranking students, they are inadequate in assessing student learning and development. The relevant literature consistently identifies the following reasons:

Many types of student ability are not captured by a standardized test.

Qualities such as a student's sense of citizenship, ethics, confidence/self-esteem, aesthetic appreciation, respect for others, self-discipline, social competence, and desire to learn, although not easily quantifiable, contribute enormously to the multiple goals of schooling. By forcing schools and teachers to concentrate on basic, quantifiable skills, standardized testing poses a serious threat to this diversity of goals and content. Medina and Neill tell us that "[t]he ability of standardized tests to accurately report students' knowledge, abilities, or skills is limited by assumptions that these attributes can be isolated, sorted to fit on a linear scale, and reported in the form of a single score." (p. 12)

In his seminal work *The Mismeasure of Man*, Gould debunked two major myths about human intelligence (as cited in Darling-Hammond, 1994, pp. 10-11): that it is an "innate, unitary, measurable commodity" and, flowing from the notion that intelligence could be quantified in this way, that individuals can be compared and ranked on this basis. Different perspectives on the concept of intelligence are emerging. Gardner (1999) notes that areas of debate include "what intelligence is, how (and whether) it should be measured, and which values should be invoked in considerations of the human intellect." (p. 70) He proposes that we all "possess at least eight intelligences: linguistic and logical-mathematical (the two most prized in school and the ones central to success on standard intelligence tests), musical, spatial, bodily-kinesthetic, naturalist, interpersonal, and intrapersonal" and that "no two of us exhibit the

same intelligences in precisely the same proportions." (p. 71) This of course has important implications for educational assessment and evaluation practices.

Tests may be standardized, but students are not.

A student's capacity to learn is affected by a host of factors including: the impact of poverty; parents' educational level; mental, physical and emotional health; the effects of racial and other forms of discrimination; and language of origin. For this reason, the absence of information on the characteristics of students in the schools taking part in testing programs leads to misleading and inappropriate use of the data derived (Canadian Teachers' Federation, 1995b). In a 1994 survey of province-wide assessment programs carried out by the Canadian Teachers' Federation, a few jurisdictions reported collecting some background data on students in conjunction with the tests (principally demographic in nature). One province collected information on such factors as parental involvement in education and amount of homework. However, no information on the socio-economic origin of students was collected in any jurisdiction (Gilliss, 1994, p. 5).

Gilliss (1993) states that "[a]nyone who expects students who are disadvantaged, impoverished, undernourished, neglected or severely disabled to score at the same level as students who are affluent, healthy, well cared for or able [bodied] is doomed to disappointment." She also observes that average test scores for schools provide more information about the pooled characteristics of individual students than about the schools themselves. Rather than incur the expense of developing, administering, grading, and ranking test scores to determine which schools score the highest, a socio-economic analysis of census data of the communities where the schools are located would provide much the same information. McLaughlin (1991) echoes this point. Responding to the call for standardized testing in "America 2000" (the national education strategy for the U.S. announced under former President George Bush), she notes that testing

> demands accountability for outputs while ignoring society's accountability for inputs. Consequently, even the best of testing strategies can do little more than document the substantial inequalities that exist throughout the U.S. school system and assume away the fundamental needs for adequate resources to support the endeavor. (p. 250)

Standardized tests designed for large numbers of students are of necessity very general in nature.

This leads to a mismatch between what is taught (given curriculum diversity) and what is tested (Casas & Meaghan, 1995, pp. 18-19). In his analysis of the Grade 8 results from the Third International Mathematics and Science Study (TIMSS) released in November 1996, Ireland (1997b) points out the substantial differences in test-curriculum match which exist among the provinces. For example, the TIMSS report noted that while 98% of the test items matched the British Columbia science curriculum, the corresponding figures for Ontario, Alberta, and Newfoundland were 53%, 54%, and 54% respectively. There was however little explanatory information on test-curriculum match beyond the tables and their respective figures. Ireland suggests that the TIMSS study could have been improved by addressing the following questions (excerpted from Ireland, 1997b, p. 14): How well did students do on questions derived from the curriculum they should have been taught? How representative of a jurisdiction's curriculum was the test and how significant is the part of the curriculum not tested? What outcomes were on the test which are not in a jurisdiction's curriculum, and how significant are those outcomes?

Standardized tests typically measure lower-order recall of facts and skills, and penalize higher-order thinking.

Meaghan and Casas (1995d) note that standardized tests "measure a student's ability to recall facts, define words, perform routine operations, rather than higher learning skills such as analyzing, synthesizing, forming hypotheses and exploring alternative ways of

solving problems". In addition, they "exclude the process of feedback and debate which is at the core of learning .... [and] they rule out independent creativity and criticism by eliminating the need to construct answers" (p. 47). The following analysis by McMurtry (1992) of a question from an Alberta Ministry of Education achievement test for Grade 3 social studies underscores this problem:

> Question 7, one of 9 Model Questions, asks "Which picture BEST shows people keeping a custom?" Underneath are four block pictures – a pick-up truck, three Native people in traditional costume, kids in uniform playing hockey, and a group of teenagers dancing. The last three of these pictures all qualify as representing "people keeping a custom." Yet only one picture qualifies as "correct." This means that a grade 3 child could be marked wrong for choosing correctly in two of the three cases, that any reason or argument over what is in fact true is ruled out even if it is fully justified, that a genuinely critical intelligence will be thrown off by a test that is demonstrably arbitrary in its selection of what is right and wrong (p. 95).

In U.S. schools, the increasing emphasis on the use of "test-oriented accountability measures" parallels a decline in teaching methods such as essay writing, research projects, lab work, and student-centred discussions designed to foster higher-order learning skills (Darling-Hammond, 1991, p. 222).

Because standardized tests are designed to sort individuals into groups, test questions are chosen on the basis of how well they contribute to spreading out the scores, not on their centrality to the curriculum or their predictive validity.

In norm-referenced tests, score dispersion is required as a basis for making comparisons. Questions that don't contribute to spreading out the scores are therefore avoided. Robertson (1998b) explains this rather technical concept in terms understandable to the lay person:

> Standardized tests actually squeeze out what is most commonly taught and learned, not because of any nefarious plot, but because of the statistical requirements of data analysis. The type of question that best discriminates among learners is the one that is answered correctly only 50 per cent of the time. Questions that are consistently answered correctly by more than 60 per cent of students are routinely dropped. While these questions may be too easy, they may also represent a core area of content that has been carefully taught to almost every child. To function as they are intended, standardized tests require the evaluation of peripheral concepts, not the common core. Forced to teach to the test, teachers must focus on less important content – the kind of factoids that suit multiple-guess exams – rather than on core ideas or complex understandings. (p. 71)

Test performance is shaped by individual characteristics not related to content knowledge.

Pressure to increase scores on standardized tests and "a lack of understanding of the complexities of achievement and its causes" have resulted in what is known as test score pollution, described as "factors affecting the truthfulness of a test score interpretation." (Haladyna, Nolen & Haas, 1991, p. 4) Test score pollution encompasses a number of test preparation and administration practices designed to increase test scores but which are unrelated to the material covered. Haladyna, Nolen and Haas (1991) describe some of these practices and whether they are ethical. They include:

- √ teaching test-taking skills;
- √ developing a curriculum and teaching objectives to match the test;
- √ presenting items similar to those appearing on the test;
- √ presenting the actual test items before the test;

- √ using commercial materials specifically designed to improve test performance;
- √ dismissing low-achieving students on test days; and
- √ interfering with responses by, for example, giving hints or answers to students or altering response sheets.

Other factors that can influence test scores include students' anxiety and motivation levels (Haladyna, Nolen & Haas, 1991). It has been demonstrated that anxious test-takers don't perform as well as test-takers who are more relaxed, and they do particularly poorly on more difficult test items. A meta-analysis of over 500 studies by Hembree (as cited in FairTest, 1995) reveals "that students with high test anxiety hold themselves in lower esteem than do those who are less test anxious .... [f]emales have higher test anxiety than males, Blacks in elementary school have higher test anxiety than Whites, and Hispanics have higher test anxiety than Whites at all ages." (p. 9)

Meaghan and Casas (1995d) note that because standardized test scores "are very sensitive to motivational factors ... [they] are not good measures of knowledge" (p. 47). Motivational factors were the focus of a study of student performance on the mathematics section of the Iowa Test of Basic Skills. In this study over 400 elementary students were randomly placed in two groups – those in the experimental group, which received special instructions from their teachers on the importance of doing their best on the test, had significantly higher scores than a control group which did not receive the instructions (Meaghan & Casas, 1995c, pp. 42-43).

Test preparation and administration take up valuable classroom time that could be used for teaching.

In some instances, curriculum is discarded to make way for test preparations that sometimes begin months before the actual test (Medina & Neill, p. 26). A number of Ontario teachers preparing students to take the Ontario Grade 9 Reading and Writing Test noted privately that much of the latter part of the 1993/94 school year was

spent in test preparation activities (Meaghan & Casas, 1995b, p. 83). In 1997 Ontario's testing agency, the Education Quality and Accountability Office (EQAO), conducted a provincial assessment of all Ontario Grade 3 students in the areas of language arts and mathematics (this annual province-wide test has since been expanded to include all Grade 6 students). Some of the limitations of this testing exercise which took ten full days to administer at a cost of over $7 million are documented in Robertson (1998b) and Ireland (1997a). While the administration time for the Grade 3 test was reduced the following year, it still required three hours per day over five days. Such exercises raise ethical questions regarding the use of valuable teaching and learning time being taken up with testing activities of limited utility.

Teachers are induced to teach to the tests rather than for learning with the result that curriculum is becoming increasingly test-driven (Meaghan & Casas, 1995d, p. 47).

This phenomenon is well-documented in the research literature. Many teachers have reported that they align their curriculum with what will be tested, are concerned that standardized testing narrows curriculum, and feel pressure to improve test scores (Meaghan & Casas, 1995c, pp. 38-39). For example, a national survey of teachers conducted by the Boston College Center for the Study of Testing, Evaluation, and Educational Policy reported that teachers with a large percentage of minority students in their classrooms (over 60%) were more likely to teach to the standardized tests, emphasizing basic academic skills and time spent in test-taking activities, than were their counterparts who taught mainly White students (Madaus, 1994, pp. 83-84). Wideen et al. (1997) examined the impact of Grade 12 school-leaving examinations on the teaching of science in several British Columbia school districts. They concluded that the high-stakes tests reduced the variety of instructional approaches used, particularly at the Grade 12 level, because they "discouraged teachers from using strategies which promoted enquiry and active student learning", and that "this impoverishment affected the language of classroom discourse." (p. 428)

Other problems

Additional problems associated with over-reliance on standardized tests include:

- √ neglecting essential educational services by allocating scarce financial resources to expensive (and questionable) standardized testing programs at a time of severe government cutbacks to education (Meaghan & Casas, 1995b, p. 87); the testing of all Ontario Grade 3 and Grade 6 students in 1999 is expected to cost about $12 million;

- √ disregarding variations in curriculum and curriculum organization for students in particular grades among different countries when comparing international test results (Ireland, 1995);

- √ the inability of information provided by standardized tests to assist educators in identifying and improving ineffective school programs (Casas & Meaghan, 1995, p. 19);

- √ shifting responsibility over curriculum from teachers and school authorities to the testing industry and government officials, thereby lessening rather than enhancing accountability (Meaghan & Casas, 1995d, p. 47).

In the United States, the production of standardized tests is mainly in the hands of commercial publishers and non-school agencies (Darling-Hammond, 1991, p. 220). The situation in Canada is similar – standardized tests are developed and marketed by the major textbook publishers with profits going to the U.S.-based parent companies (Barlow & Robertson, 1994, p. 120). The U.S. private testing industry is a billion dollar a year business which, despite its growing influence on the education system, is virtually unregulated and unaccountable. Under such conditions, increased involvement by teachers and parents in the educational decision-making process is ruled out (Medina & Neill, 1990).

Testing and Inequity in Education

Not only is there little evidence to support standardized testing from a pedagogical standpoint, it is clear that standardized testing can have a detrimental impact on the most vulnerable students. Far from being a neutral practice, it perpetuates and intensifies educational inequities in two ways: through the misuse of test scores; and because test bias works against the interests of students from low-income groups, racial and ethnic minorities, girls and young women, and students with disabilities. Darling-Hammond (1994) reminds us that "[t]he role of testing in reinforcing and extending social inequalities in educational opportunities has by now been extensively researched ... and widely acknowledged." (p. 10)

Inappropriate use of tests

Tests are often used inappropriately. The history of testing reveals many instances of tests developed for a specific purpose being used for other unintended purposes. The Scholastic Aptitude Test or SAT is a prime example. Originally designed to predict a student's potential for success in their first year of college in the U.S., the SAT has been used as a criterion for determining such things as athletic eligibility, student loan eligibility and awarding scholarships (Madaus, 1991, p. 227). Despite its limitations (including its "limited predictive value"), Schwartz (1999) observes that the SAT "has arguably become the single most important test for American high-school students – an academic and psychic rite of passage that strongly influences future educational options, prompts fierce anxiety and serves as an almost mystical barometer of self-worth." (p. 30)

Tests of individuals are used to analyze policy, program, school and teacher success, and they are being inappropriately used as "educational gatekeepers" to make important decisions about students, teachers, schools, and the school system as a whole (Medina & Neill, p. 24). Standardized test scores are becoming the mechanism that facilitates a number of questionable education practices that contribute to education inequity.

Tracking

In the U.S., test scores are used routinely to make important decisions about student placement. The "tracking" that results streams children from disadvantaged socio-economic backgrounds into average and low-track classes and programs. Most progressive educators conclude that the practice of tracking fosters a negative self-image and creates reduced expectations for a student's future. It also denies those in the lower tracks the benefits of working with higher achievers. In Canada there is a danger that standardized testing will be used as a mechanism to renew interest in tracking thereby moving us away from a more inclusive model of education (Meaghan & Casas, 1995a). Tracking accounts in large part for the disparity in achievement observed between American White students and those from minority groups, as well as between students from higher- and lower-income groups (Darling-Hammond, 1991, p. 222). Perhaps because curriculum in "low-track" classes tends to be limited and rote-oriented, students in these classes don't do as well as their peers of comparable aptitude in academic or untracked programs. Oakes (1985) concludes that tracking hurts low-income and minority group children, and that it not only reflects but perpetuates class and racial inequalities in the larger society. Using a method of testing already biased against certain groups of students to subsequently assign them to lower track programs only adds insult to injury.

Grade promotion and retention

Standardized test results have also been used to determine grade promotion and retention. The practice of grade retention can contribute to an increased drop-out rate (Medina & Neill, p. 29) and

according to Shepard and Smith, " '[c]ontrary to popular beliefs, repeating a grade does *not* help students gain ground academically and has a negative impact on social adjustment and self-esteem.' " (as quoted in Darling-Hammond, 1991, p. 222) Grade retention has also resulted in lower achievement. Given the lower levels of self-esteem generally observed among girls as compared to boys, the implications of this policy for further eroding self-esteem among girls and young women requires serious consideration.

Graduation

Test scores have been used to determine a student's eligibility for high school graduation. According to Darling-Hammond (1994),

> [t]he rationale for this practice is that students should show that they have mastered the "minimum skills" needed for employment or future education in order to graduate. The assumption is that tests can adequately capture whatever those skills are. While this appears plausible in theory, it is unlikely in reality, given the disjunction between multiple-choice tests of decontextualized bits of information and the demands of real jobs and adult tasks .... *In fact, research indicates that neither employability nor earnings are significantly affected by students' scores on basic skills tests, while chances of employment and welfare dependency are tightly linked to graduation from high school* [italics added] .... Thus, the use of tests as a sole determinant of graduation imposes heavy personal and societal costs, without obvious social benefits. (p. 14)

Nineteen U.S. states currently require students to pass state tests to graduate from high school (Education Week on the Web, 1999). In Canada the majority of provinces now require graduating students to take high school exit exams with the stakes ranging from 30 to 50 per cent of the final course grade (Dunning, 1997). Once taken as evidence of years of cumulative development, the high school diploma may eventually come to be replaced by a one-shot, high-stakes test.

Other manifestations of test misuse

Misuse of the results of province-wide assessment programs in Canada includes "inadequate representation of test results by politicians and the media and inadequate explanation of test results to the public (for example, no background information or comparison of samples with differing characteristics)." (Gilliss 1994, p. 6) A survey of provincial ministries and departments of education as well as a sample of local school boards across the country also reported that a major drawback of province-wide testing programs is the misinterpretation of results by the public and the media (Traub, 1994, pp. 1, 3).

When 90% of the students in an Ontario Grade 9 reading and writing test fell into the category of having "adequate" reading skills, results that received broad coverage in the media, the North York School Board vice-chair "announced that these results were not good enough for his Board" – this despite the fact that

> he had not been involved at any level in the preparation of the test, had no indication of how reliable the scores were, or whether the student population in North York [in Metropolitan Toronto] has special characteristics such as a higher-than-average proportion of ESL [English as a second language] students. (Meaghan & Casas, 1995b, p. 83)

Standardized test results are used in some U.S. jurisdictions to make decisions about teachers with respect to merit pay awards as well as certification and recertification. Test-based decisions impacting on schools and school systems include such things as allocation of funds, and school system certification and decertification (Medina & Neill, p. 24). To use average school test scores as the basis for allocating rewards and sanctions to teachers and schools, a practice several U.S. states and school districts have already put in place, is to endorse a market-based approach to stimulating improvement in the school system. Former U.S. President Bush proposed using a national test to

allocate federal funds to schools in "America 2000". Such a policy would provide an incentive to schools to close their doors to students who might "bring down" the average scores, such as children with learning and other disabilities, those whose first language is not English, "or children from educationally disadvantaged environments." (Darling-Hammond, 1991, p. 223) It could also result in experienced teaching staff avoiding schools which serve disadvantaged kids, thereby closing off further educational opportunities for these children.

Citing a study of a large urban American school district that used average school test scores to make decisions about rewards and sanctions for individual schools, Darling-Hammond (1991) ominously describes the impact as "the widespread engineering of student populations". The authors of the study state that

> "[s]tudent selection provides the greatest leverage in the short-term accountability game .... The easiest way to improve one's chances of winning is 1) to add some highly likely students and 2) to drop some unlikely students, while simply hanging on to those in the middle. School admissions is a central thread in the accountability fabric." (as quoted in Darling-Hammond, 1991, p. 223)

As noted above, the pressure to raise average test scores can result in the rejection by the school system of more at-risk students. This includes placing students with low scores in special education programs to avoid having their scores reflected in school reports; refusing to admit students to "open enrolment schools" on the basis of low scores; and going so far as to encourage students with low test scores to drop out of school (Darling-Hammond, 1991, p. 223). In addition, the pressure to raise and maintain average test scores is such that,

> as teachers, principals, and superintendents come to see test scores as having a critical impact on their own careers and on their students' lives, testing becomes the curriculum. Dissemination of test scores places greater pressure on teachers and administrators, who respond to the higher stakes by devising ingenious ways to raise

> scores without necessarily upgrading skill levels. (Meaghan & Casas, 1995c, p. 49)

As Elliot Eisner remarked in a presentation at the B.C. Teachers' Federation AGM in March 1998, "[i]t is perfectly possible for a school's faculty to raise test scores and at the same time to diminish the quality of education."

High-stakes, measurement-driven accountability

A recent *Education Week* survey of the accountability policies of 50 states (with the less than subtle title "Quality Counts '99: Rewarding Results, Punishing Failure") reveals the strong push for outcome-based accountability to leverage educational reform in the U.S. According to the report, among the components (described as "essential steps") necessary for putting a comprehensive system of accountability in place are tests, standards, report cards, rating systems, and rewards and sanctions. Using these components as the basis for its analysis, the survey found that:

- √ 48 states have state-wide testing programs (13 of which use only multiple choice tests).
- √ 36 states publish annual "report cards" on the performance of individual schools.
- √ 19 states evaluate and issue public ratings of school performance including "low-performing" schools.
- √ 16 states have the power to close, take over or "reconstitute" a failing school (reconstitution is described in the report as "replacing all or most of a school's staff and reopening it under new management.") Not surprisingly, schools attended by poor and minority students are more likely to have sanctions applied to them.

- √ 14 states provide monetary rewards for individual schools based on performance.

- √ 2 states have attempted to link the evaluation of individual teachers to student performance.

The report contrasts this "hard approach to accountability" with what it describes as a "soft approach" – the latter involves providing schools and students with sufficient resources, support and encouragement as well as enhancing professional accountability. The concept of professional accountability recognizes the critical link between teacher quality and student achievement and hence, the importance of investing in teacher professional development and growth.

In March the Alberta government announced its own version of outcome-based accountability for K-12 education. According to an Alberta Education news release (March 11, 1999), it intends to spend $66 million for a School Performance Incentive Program (SPIP) described as an "innovative initiative designed to improve student learning" by awarding additional funding to school boards "based on how well they achieve improvement targets." Targets will be set by the province in conjunction with local boards. Allocation of funding awards (ranging from 2-4 per cent of a board's salary budget) is conditional on improvements in learning as measured by provincial achievement test and diploma exam results, high school graduation rates, and "locally determined performance indicators" such as student attendance and results from local achievement tests. The program which is voluntary (and apparently unprecedented in Canada) will be implemented in the 2000/2001 school year. As noted above, Americans are more familiar with such initiatives – 14 states including North Carolina, Maryland and Texas have instituted some form of financial incentive program ("Alberta Proposes Bonuses for Student Achievement Gains", 1999).

The Alberta Teachers' Association (ATA) is concerned that a program which essentially offers cash bonuses in return for improved test scores, particularly at a time of budget cutbacks, "will pit schools and boards against each other and unfairly favor well-to-do districts." (McMahen,

BEST COPY AVAILABLE

1999) Other questions raised by the SPIP include: What will be the impact on teacher collegiality? Will the program result in increased teacher requests to transfer to schools with higher student achievement? What is the potential fallout from identifying schools and classrooms that don't meet the target objectives? ("Alberta Proposes Bonuses for Student Achievement Gains", 1999) Arguing that such incentive programs are inconsistent with what is known about effective teaching and student evaluation practices, teacher motivation and school improvement generally, the ATA and its provincial education partners have proposed an alternative to the SPIP designed to improve teaching and learning based on proven research and practice. The Minister of Education has agreed to put the SPIP on hold pending discussion of the alternative program ("Improvement, Not Incentives", 1999).

The so-called hard approach as described in the *Education Week* survey is clearly not the path to genuine accountability in education as advocates suggest. Darling-Hammond (1994) explains the negative impact on equity of such an approach:

> Needless to say, this kind of policy that rewards or punishes schools for aggregate test scores creates a distorted view of accountability, in which beating the numbers by playing shell games with student placements overwhelms efforts to serve students' educational needs well .... Applying sanctions to schools with lower test score performance penalizes already disadvantaged students twice over: having given them inadequate schools to begin with, society will now punish them again for failing to perform as well as other students attending schools with greater resources and more capable teachers. This kind of reward system confuses the quality of education offered by schools with the needs of the students they enroll; it works against equity and integration, and against any possibilities for fair and open school choice, by discouraging good schools from opening their doors to educationally needy students. Such a reward structure

> places more emphasis on score manipulations and student assignments or exclusions than on school improvement and the development of more effective teaching practices. (pp. 15-16)

In the context of U.S. proposals to introduce national testing, Madaus and Kellaghan (1993) analyze the use of testing as an instrument to drive educational policy noting that "the link between test performance and important rewards or sanctions is the key ingredient that gives tests their policy punch" – hence the term "high-stakes testing". In an article entitled "Testing as a Mechanism of Public Policy: A Brief History and Description", they outline six basic principles to describe the processes "through which tests used in the policy sphere affect individuals and institutions". Their observations are insightful. (Excerpted from Madaus & Kellaghan, 1993, pp. 8-9):

1. The power of tests and examinations to affect individuals, institutions, curriculum, or instruction is a perceptual phenomenon – in other words, if students, teachers, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false – the effect is produced by what individuals perceive to be the case.

2. If important decisions are presumed to be related to test results, then teachers will teach to the test.

3. In every setting where a high-stakes test operates, a tradition of past exam questions develops, and this tradition eventually comes to define de facto the curriculum.

4. When test results are the sole or even partial arbiter of future educational or life choices for students, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indicator of achievement.

5. A high-stakes test transfers control over the curriculum to the agency that sets or controls the exam.

6. There is a very general principle that is true of all important social indicators, including tests, and which relates to a process that is probably always unintended and very often not anticipated. The principle embodies the proposition that the more any quantitative social indicator is used for social decision making, the more likely it will be to distort and corrupt the social processes it is intended to monitor.

Authentic forms of student assessment as an alternative to standardized testing are not immune to these effects. There are some indications that when authentic assessment (which will be discussed in more detail in a subsequent section of this report) is tied to high-stakes accountability, the instructional benefits of such assessment are undermined (Jones & Whitford, 1997; Torrance, 1993). For example, Kentucky's high-stakes accountability system converts the results of performance-based assessments into a "school score" and relies heavily on the use of external rewards and sanctions. According to Jones and Whitford (1997),

> [i]f the score exceeds the state's expectation for a school, the teachers and administrators can receive substantial bonuses. If the score is not high enough or does not continue to improve over time, the teachers and administrators can be placed on probation, and the school can be taken over by the state. (p. 276)

Particularly frustrating for teachers and administrators was the state's decision, lacking empirical data, to design the formula for determining school target scores (as a means of measuring the rate of school improvement) in an arbitrary manner.

One of the effects of KIRIS (Kentucky Instructional Results Information System) has been increased state control over curriculum at the expense of local control and the professional judgement of teachers. As Jones and Whitford explain,

> [p]ressure generated by the state test for high-stakes accountability has led school-based educators to pressure the state to be more explicit about content that will be tested. This in turn constrains local school decision making about curriculum. This dialectical process works to increase the state control of local curriculum. (p. 278)

Rather than generating constructivist, performance-based classrooms as was perhaps the intent, KIRIS has instead emphasized "learning that can be measured reliably and validly" – the testing of observable, measurable outcomes effectively determines what is worth knowing. While performance assessment holds promise for improving student learning, this potential is corrupted in a high-stakes environment. The authors conclude that the principles underlying high-stakes accountability and performance assessment are fundamentally incompatible: the former "encourages conformity to externally imposed standards, while the [latter] grows out of emergent interactions between teachers and students in dynamic local classrooms." (Jones & Whitford, p. 280)

Bias in testing

Standardized tests are typically gender-biased.[5] Rosser et al. (1992) observe that while young women consistently obtain higher grades in *both* high school and college in the U.S., they receive lower scores on

[5] Robertson (1992) observes that "differences between male and female achievement do not necessarily demonstrate bias. It should be remembered that one group may genuinely know more about a particular content area than another group. Any bias is thus the result of the kind of knowledge which is valued, rather than bias in favour of or against a particular group. However, if a test fails to meet the criterion of validity, that is, if two groups know the same amount about a subject but one group consistently scores higher on a test alleged to evaluate that knowledge, or if a test fails to identify the superior level of knowledge of one group over the other, then such a test can be said to be biased." (p. 26)

the Scholastic Aptitude Test (SAT) and similar tests which allegedly predict students' future academic performance.[6] High school girls aspiring to attend college represent over 52% of those writing these tests. However, their abilities are systematically underestimated by these exams resulting in reduced opportunities for attending post-secondary institutions, for obtaining scholarship funding, and for acceptance to special programs for gifted students. Students who receive low SAT scores have reduced expectations of their abilities which in turn affects their decisions about which colleges they apply to (Weiss, 1991, p. 108). The longer term economic and personal consequences for women include fewer good employment opportunities, lower income levels and reduced self-confidence. Female students from low-income backgrounds face the combined gender and class biases of the exams and are effectively penalized twice (Rosser et al., 1992). In short, SAT scores tell us more about factors such as family income, race and gender than they do about one's ability to perform well in university (Robertson, 1998b, p. 80).

Robertson (1992) discusses the impact of testing on girls in the context of math, science and technology (MST) education. She notes that "complex indicators of 'success' for students and for education are frequently and inappropriately reduced to single-test numbers" (p. 25) and that there is clearly an important political agenda underlying this direction in education reform. For high school students, small differences in test score results not only decide whether one passes a

[6] Even the world's largest private educational test maker, the Educational Testing Service (ETS) which develops and administers the SAT, has admitted the SAT's bias in terms of over-predicting male college performance while under-predicting the college performance of females. ETS also acknowledges that high school grades are a better predictor of college success than their tests (Thurmond, 1994). In a landmark U.S. federal court decision on the discriminatory nature of the SAT, the judge stated that " '...[T]he probability, absent discriminatory causes, that women would consistently score 60 points less on the SAT than men is nearly zero.' " Even after accounting for such factors as ethnicity and parental education, a substantial gap in points remained (Rosser et al., 1992, pp. i-ii).

course but determine which students will be accepted to select universities (p. 25). As "scarcity" of post-secondary opportunities flows from underfunding of the post-secondary education system, Canadian universities have increased tuition fees as well as reduced the number of spaces available. These factors combined will reduce access to higher education for women, particularly in areas (such as MST) in which they are already under-represented.

The following findings (excerpted from Robertson, 1992, p. 26) are taken from the American Association of University Women (AAUW) report *How Schools Shortchange Girls* (1992). This important study synthesizes the available research on the educational experience of girls concluding that the education they receive is inequitable in terms of quality and quantity to that received by boys:

- √ Despite girls' receiving higher grades in high school and college, boys were twice as likely as girls to have received a college scholarship based on the results of standardized high school tests.

- √ Girls have consistently outscored boys on the SAT-verbal when test content refers to concepts and ideas rather than things, or to aesthetics, philosophy, relationships or stereotypically female tasks; whereas boys have scored better on items relating to science or sports; since the SAT-verbal was revised in 1978 to increase the number of science-related items, reading subscores favouring boys have increased from three to twelve points.

- √ As noted earlier, SAT scores under-predict the success of young women, and over-predict the success of young men in college/ university achievement, especially in math.

- √ Bias in test items can occur within content, format and the skill or knowledge area being tested; for example, in the area of math, girls tend to score higher on computational tasks, algebra and logic, whereas boys tend to score higher on problem-solving and geometry – tests which give more weight to one cluster of skills over another will produce gender differences.

- √ Girls tend to score higher on essay and open-ended items; boys tend to score higher on multiple choice items; girls complete fewer test items than boys, are more likely to not finish tests, and more likely to choose "I don't know" as an answer when it is offered.

FairTest notes that

> [t]he timed, speeded nature of the [SAT] works against young women and members of minority groups. Guessing is also a barrier for females, and possibly for people of colour. The forced-choice format does not allow for shades of meaning, working against girls' more complex thinking-style. (Rosser et al., 1992, Appendix I)

In a follow-up report to *How Schools Shortchange Girls* entitled *Gender Gaps: Where Schools Still Fail Our Children* which looks at progress toward equity since the early 1990's, the AAUW Educational Foundation (1998) finds that "high-stakes tests, with disproportionate power to affect students' lives, are the tests that most dramatically reflect gender differences in performance." (p. 35) Among the report's recommendations are that post-secondary institutions formulate decisions about student admissions and scholarship eligibility using a "broad range of material"; and that research focus on the relationship between student test scores and grades in the context of gender, race and ethnicity in order to provide information "to develop appropriate equity responses based on individual differences." (p. 41)

Robertson states that although the AAUW and other data can't be used as evidence that gender bias is disadvantaging Canadian girls in every MST-related evaluation, she says "it is curious to note how little attention is paid to the apparent risk to young women posed by a massive shift to computer-scored, single-right-answer standardized evaluation." (p. 27)

Most standardized tests in the U.S. are designed for, and written by, a middle- to upper-class White population and as a result reflect a particular perspective. Studies have shown that a number of characteristics of standardized tests can bias the results against minority students and students from low-income backgrounds. These include the following (from Medina & Neill, pp. 8-10):

- √ Language bias manifests itself in English-language tests taken by students whose first language is not English; it manifests itself more subtly in tests written in a more formal, stylized version of English, rather than the common vernacular, for students who speak non-standard English dialects.

- √ The content of test questions seeks responses which ignore the cultural experiences, perspectives and knowledge of children from racial and ethnic minorities, low-income families, and inner city and rural children.

- √ Individuals have "different ways of knowing and problem-solving". These different learning styles (not different abilities), often associated with such factors as race/ethnicity, income level and gender, are not considered in test design, the assumption being "that all individuals perceive information and solve problems in the same way."

In their research on language testing of Black South African students, Peirce and Stein (1995) found that students' multiple interpretations of a text for a reading test were socially constructed. Their findings have broad implications beyond the South African context in terms of education inequity. They acknowledge that,

> [i]n essence, if test makers are drawn from a particular class, a particular race, and a particular gender, then test takers who share these characteristics will be at an advantage relative to other test takers .... To promote equity in educational assessment, different stakeholders, such

> as testers, teachers, administrators, parents, and students should be able to contribute to the test development process. (p. 62)

The injustice of being denied access to post-secondary studies because of scoring low on a biased test is underscored in the following passage by Weiss (1991):

> For those minority students who have overcome innumerable social and economic barriers and achieved academic success in early schooling, the denial of access to higher education because of low college or graduate school admissions tests is particularly frustrating. (p. 114)

Coaching courses to prepare students to take tests such as the SAT have evolved into a multi-million dollar industry. The largest SAT preparation companies are the Princeton Review and Kaplan Educational Centers with the latter generating annual revenues approaching $300 million (Schwartz, 1999, p. 32). Weiss notes that a coaching course can increase one's SAT score by enough to "mean the difference between a rejection notice and college admission with a scholarship." (p. 118) Such courses however can be expensive – a 6-8 week course including practice tests costs about $800 while private tutors can charge a few hundred dollars per hour (Schwartz, p. 32) – putting them well beyond the means of the majority. This places minorities and low-income students at a double disadvantage – according to Weiss, "[n]ot only are they unable to afford the advantages promised by coaching, but the success of coaching increases the disparity between racial groups even further." (p. 118)

Testing has become an entrenched practice in our society, not only in education but in other spheres.[7] Some have described testing as a form of social technology arguing that, although there is nothing inherently unjust about technology itself, "technologies are a product of the existing structure of opportunities and constraints in society" and as such, "they are likely to extend, shape, rework, or reproduce this structure", and that includes any inequities built into the structure (Madaus, 1994, pp. 77, 79). In other words, the values, biases and assumptions of the elite group who create technology are reproduced in those technologies. Some of the underlying values that inform testing are utilitarianism, economic competitiveness, technological optimism, objectivity, bureaucratic control and accountability, numerical precision, efficiency, standardization, and conformity. In this way says Madaus (1994), "testing not only shapes important educational values, such as conceptions of educational quality, and student- and teacher-reward systems, but also the content and method of what should be taught and learned." (p. 79)

[7] In Testing Testing: Social Consequences of the Examined Life, F. Allan Hanson examines the social and cultural assumptions behind current U.S. testing practices. He analyzes testing as a social institution, one so closely associated with the notion of meritocracy that we rarely ask the difficult questions such as "why test at all?" Drawing on the work of Foucault, he proposes that testing (from lie detection and drug use tests to vocational, intelligence and achievement tests) constructs who we are as social beings, and that tests serve as a form of surveillance and control over the individual by organizations and institutions that use and value these tests (Kleinsasser, 1994, pp. 96-97).

A Few Words About Authentic Assessment

Authentic or performance assessment is often proposed as an alternative to standardized testing. FairTest describes performance assessment as a "general term for an assessment activity in which students construct responses, create products, or perform demonstrations to provide evidence of their knowledge and skills" (Neill et al., 1997, Appendix B). Such assessment covers a wide range of initiatives including: oral presentations, debates, exhibitions, collections of students' written work, videotapes of performances and other learning experiences, constructions and models and their solutions to problems, experiments and the results of scientific and other inquiries, teacher observations and inventories of individual students' work and behaviour, and cooperative group work (Darling-Hammond, 1994, pp. 5-6).[8] Unresolved issues related to this relatively new form of assessment include the substantial time commitment required for development and administration (authentic assessment

[8] In her review of F. Allan Hanson's book Testing Testing: Social Consequences of the Examined Life, Kleinsasser (1994) notes that the results of standardized testing, which, given the limitations of such testing, provide only indirect evidence of a student's academic achievement, are retained by school authorities as part of the student's official school record. This evidence is then used to make important education-related decisions affecting the student's future; whereas the actual work (or direct evidence) produced by a student, whether writing samples like essays or stories, art, science fair projects, audio- and video-tapes of music and dance performance, etc., is usually kept by the student (rather than the school) because it is more valued by its creator. These products however "are much less meaningful in the hands of powerful authorities like schools or employers .... When institutions choose to prize test results that are held by authorities with the power to recommend or not recommend, they send a strong message to test takers about society's values and evaluation criteria." (p. 98)

is by nature time- and labour-intensive) and the associated financial costs as well as the trade-off between validity and reliability.

Changes to the form and content of assessment (as in the case of authentic assessment) will not by themselves solve the problems associated with more traditional assessment. Darling-Hammond (1994) stresses that assessment must also be viewed in terms of how it will be used and in particular, how it will be used to support and enhance equity initiatives within the education system. She argues that

> *changes in the forms of assessment are unlikely to enhance equity unless we change the ways in which assessments are used as well*: from sorting mechanisms to diagnostic supports; from external monitors of performance to locally generated tools for inquiring deeply into teaching and learning; and from purveyors of sanctions for those already underserved to levers for equalizing resources and enhancing learning opportunities. [italics in original] (p. 7)

At the same time, broader reforms must be implemented including adequate funding and supports for social, health, family and educational resources, what Madaus (1994) refers to as creating "a level playing field for students and schools" in terms of social and educational conditions. He states that

> [e]quity issues arising from assessment are not confined to the specific instrumentation, technique, or psychometrics inherent in the test itself. More broadly, equity issues relate to antecedent and current social and educational conditions. (p. 87)

As with any new technology and the aura that envelops it, there is a danger that authentic assessment may come to be viewed as *the* solution to educational problems, to the exclusion of addressing other factors such as student health and living conditions, teacher training, etc. Madaus (1994) warns against having technological solutions "blind

policymakers and the public to the reality that we Americans cannot test, examine, or assess our way out of our educational problems" (p. 78), a caution we would do well to heed in the Canadian context given our growing emphasis on test-based accountability.

Politics of Testing

Standardized testing needs to be viewed with a political as well as a pedagogical eye. Advocated as a measure of educational achievement and a means of addressing our educational problems, standardized testing is politically attractive because, while creating the illusion that something is being done to improve the quality of education (via what is perceived as an objective, neutral mechanism), it allows us to ignore the complex reality of educating children and youth, and the associated demands, time and costs. McMurtry (1992) notes that standardized tests "can be a destructive distraction from what we need most to do and to achieve as educators" and that "they can be politically misused to escape the actual demands of teaching." (p. 95) In a chapter on the politics of testing in *No More Teachers, No More Books: The Commercialization of Canada's Schools*, Robertson (1998b) states that "[i]t is so much easier to increase the quantity of testing than to improve the quality of teaching – or to deal with inequality of opportunity." (p. 72) She also observes that testing begets more testing:

> Whether the alleged deficiency [with schools] is curriculum, standards, or accountability, the strategy is to use data, no matter how corrupt, to leverage education reforms. Thus a key political mechanism is to increase testing – more tests for more kids more often, producing more data, more widely reported and more susceptible to manipulation. (pp. 63-64)

International comparisons

Despite the fact that international comparisons of educational achievement are difficult to do, the available data shows that Canada is near the average for comparable countries, and in some cases fares better (Rehnby, 1996, p. 15). Nevertheless, the perception of serious

problems with our schools has been fueled in part by sensationalist and distorted media coverage of Canadian performance on international standardized tests. Using performance on international tests as an indicator of problems with our education system can be misleading. Some of the methodological problems associated with international testing of math and science achievement include the following (from Canadian Teachers' Federation, 1995a):

- √ Participation in international studies is often voluntary, and not all school systems are involved. Rankings therefore reflect only those countries who choose to participate.

- √ Participating countries have different ways of choosing the students who will write the tests (some select only their highest achieving students).

- √ Participating countries are not rated against their own curriculum, and there is no effort to adjust test scores to reflect circumstances in different countries.

Past experience with international studies has clearly demonstrated that "quantitative data can be misleading without detailed qualitative information to put it in context." (Canadian Teachers' Federation, 1995a) Nagy (1995) notes that "[i]nternational studies have the potential to develop as a useful tool for understanding and improving education, but they face a conundrum. As they become, of necessity, more sophisticated, their very complexity will make simplistic misinterpretation difficult. This, however, may result in loss of their political appeal and funding."

Public and political misinterpretation of complex international comparative studies can in turn drive the policy development process, resulting in misinformed decisions. Nagy (1996) makes an important distinction between *policy* judgements and *political* judgements – the former are informed decisions that flow from the original reports, accompanying cautions on data interpretation and any relevant academic and specialist analyses; the latter refer to uninformed announcements based on "the media and political view of these

reports." (p. 1) Unfortunately, the latter have become all too commonplace.

Misrepresentation of test results (whether at the international, national or provincial level) can heighten the concerns of parents, causing them to feel the education their children are getting is inferior to that received by children in other countries, provinces or districts. Politicians and policy makers respond to the perceived crisis of public confidence by placing an increased emphasis on standardized testing as a measure of accountability. Meaghan and Casas (1995b) state that

> testing has become the quick-fix solution for school board administrators and ministry of education officials who need to produce "results" and need to do so quickly. Tackling curricular issues, improving the quality of teacher training, [and] developing fair and efficient assessment methods would take too long to produce results and those results would not be as simple to measure and to interpret as test scores are. (p. 84)

Is there evidence to support a crisis of public confidence in schools? Not according to OISE/UT researchers Hart and Livingstone (1998). In their analysis of trends in attitudes about schools drawn from Ontario and national opinion surveys dating back to 1980, they sought to shed light on this question. Their analysis also addressed the question of whether there was "support for an agenda of fiscal austerity and privatization that is the neoconservative prescription for education" (p. 1).[9] Their research reveals sharply divergent

[9] The authors define neoconservatism generally as a "New Right political position ... which argues for: (a) a reduction in the role of the state, particularly with regard to social services, redistribution of income, and equity initiatives; (b) a reduction in government deficits and debt; and (c) a reliance on market mechanisms and competition to make public services more efficient through privatization or by bringing greater competition into the public sphere." (p. 2)

attitudes between corporate executives and the general public with respect to confidence in schools. Specifically, Hart and Livingstone found little evidence to support the notion of "an impending crisis of public or parental confidence policy preferences of the general public" (p. 16).

Testing, schools and market competition

There is also strong support from the corporate sector for national education standards and national testing. The successful lobbying efforts of organizations representing powerful business interests such as the Business Council on National Issues, the C.D. Howe Institute and the Conference Board of Canada have put standardized testing high on the government and public agendas. Standardized testing fits well with business priorities – computer-grading of multiple choice tests is cheaper and quicker than teachers marking essays and other writing assignments (Barlow & Robertson, p. 118); standardized test scores represent the educational equivalent of the corporate bottom line (Reardon, Scott & Verre, 1994, p. 2); and as noted earlier, there are large profits to be made from educational testing.

Business also supports standardized testing because it is consistent with a free market approach to education. School choice by parents acting in their capacity as consumers of education is advocated as a means of fostering competition among schools. Competition is regarded as the necessary incentive to the achievement of quality education (as "inefficient" schools close and the "good" schools rise to the "market challenge"). In this model, a basis for comparing and making decisions about schools is required: enter the ranking and reporting of average scores on standardized tests as the basis for making these decisions (Berthelot, 1995, p. 4). Of course, rank ordering schools (or districts, provinces or states) by test scores is a seriously flawed practice. Not only are scores often presented with little or no contextual information but small increases in the test scores of individual students can translate into significant changes at the state level in terms of positioning within a rank-ordered table, something which can be

manipulated to political advantage. When this fact is combined with the strong pressure to increase test scores and the questionable practices that ensue, the potential for compounding distortions and providing misleading information is very real.

In a system of school choice (which would include charter schools) the so-called "good" schools are able to attract and select the best students. The "creaming off" of motivated, high-achieving students and their more educationally-involved parents by choice schools is beginning to have a negative impact on local public schools which are left with fewer students (those that are higher-risk and higher-cost) and hence, fewer tax dollars (Dehli, 1998, p. 31).

According to the Canadian Teachers' Federation (1997),

> [s]chool-by-school competition creates high demand for the students who already have the greatest chance of success in our schools, and creates incentives to avoid serving harder-to-teach kids and kids with disabilities who may consume more than their 'share' of resources, yet not increase the school's profile in key marketing areas such as standardized test scores. (p. 17)

It should also be noted that when high-achieving students are attracted by and concentrated in choice programs, the illusion of increased achievement may well be created. Fuller and Elmore (as cited in Dehli, 1998, p. 31) caution that " 'it is unlikely that choice will do anything other than simply move high achievers around from one school to another, mistaking the effect of concentrating strong and motivated students for an effect of the school or the choice system.' "

'Pan-Canadianization' of curriculum and assessment

The growing Pan-Canadian (i.e. national) agenda for education being led by the Council of Ministers of Education, Canada (CMEC), as outlined in its 1993 "Victoria Declaration", includes major initiatives in the areas of curriculum, student assessment, and education indicators. With respect to curriculum, in 1997 the CMEC released its *Common Framework of Science Learning Outcomes: Pan-Canadian Protocol for Collaboration on School Curriculum*, a common curriculum framework which lays out the learning outcomes for K-12 science education (this project involved all provinces except Quebec). Robertson (1998b) observes that the "CMEC sidestepped the thorny constitutional problem of curriculum as a provincial matter by declaring the document to be only a framework for curriculum developers, not an actual curriculum. But no one was fooled." (p. 86) She notes that the curriculum consultation process was also sidestepped. The time available for feedback on drafts of the long document was short and the nature of the commentary was tightly prescribed – "[t]hose wishing to comment were told that they could question neither the desirability of a national curriculum nor its contentious devotion to outcomes." (p. 86) Teachers' organizations, school board associations and other groups traditionally involved in curriculum development were effectively excluded from the process.

Curriculum is also being developed regionally through the Atlantic Provinces Education Foundation in the east and the Western Canadian Protocol for Collaboration in Basic Education in the west. Such curriculum reform initiatives are in keeping with an increased emphasis on standardized outcomes and the core subject areas of math, science and language arts (Dunning, 1997).

With province-wide assessment programs in place in most provinces and territories, a full-blown national testing program cannot be far off. In her analysis of the School Achievement Indicators Program

(SAIP)[10], a national project initiated in 1989 by the CMEC to test Canadian students aged 13-16 in math, reading and writing, and science, Barker (1991) argues that the SAIP is a politically-motivated initiative developed in response to criticisms of education and calls for increased accountability from business, industry and government. She suggests that the SAIP can be viewed as a catalyst for educational reform in Canada – from more limited changes in terms of institutionalizing national testing, to sweeping reforms such as increased control over education by the federal government and the institutionalization of national education standards with economic productivity. Some of this is echoed in the 1994 report of the National Advisory Board on Science and Technology's Committee on National Standards in Education. Among its recommendations to the Prime Minister were that:

√ "the federal government assist in the development and administration of systemic tests *for all core subjects* [italics added] building on the work of the School Achievement Indicators Program", and that

√ "the frequency of the administration of the tests in all core subjects be increased to that which maximizes the information necessary to ensure continuous improvement. If maximum advantage is gained with yearly testing, then that should be instituted." (p. 23)

[10] In the first cycle of the SAIP, assessments were administered in mathematics (1993), reading and writing (1994), and science (1996). The SAIP's second cycle included assessments in mathematics in 1997, reading and writing in 1998, and science assessment in 1999. A third cycle has since received approval. Funding for the SAIP comes from the provinces, Human Resources Development Canada at the federal level, and private corporations (Investors' Group Incorporated and Spar Aerospace Limited). The CMEC estimates total costs for administration and implementation of the program from the planning stage in 1989 through to its completion in 1999 to be approximately $11 million, excluding costs incurred for provincial in-kind support [from the CMEC web site at: http://www.cmec.ca/saip/]. Although the SAIP has been in existence for the past decade, there appears to be little in the way of critical analysis of the program and its impact.

The Committee also called for national standards in education and, in a section on the need for employability skills testing, recommended that "the assessment of basic [employability] skills, once they have been defined, be incorporated into the SAIP as it develops and expands its mandate." (p. 23)

Christensen's (1998) analysis of the CMEC, from its origins in the Canadian Education Association to its emergence as a self-described "national forum on educational policy", documents the growing influence of an organization which she notes "has largely escaped critical inquiry since its birth thirty years ago" (p. iii), is not accountable to any government and, despite its stated commitment to consult and cooperate with other educational organizations, is "exclusionary in nature". In a chapter on the CMEC's work in curriculum and student assessment (whose early initiatives in these areas in the 1980's signalled the start of the CMEC's rising influence), the author describes how federal government pressure in the early part of the decade "precipitated the Council's hasty implementation of SAIP in 1993." (p. 83) Federal interest in addressing "public discontent about the educational 'crisis' in Canada" was articulated in a 1991 throne speech in which the Prime Minister referred to the need for national education goals, marking as Christensen notes "the return of the federal government to discussions about 'national' education polices" (p. 83). (Interestingly, at about the same time the U.S. government released its America 2000 education strategy calling for a system of national standards and assessment.) Mindful not to encroach on provincial jurisdiction, the vehicle chosen by Ottawa to address education issues was the Canadian Forum on Learning, set up through the Steering Group on Prosperity (Prosperity Initiative), to " 'lead a process to develop a statement of expectations for education and training in Canada' " (p. 84). At a meeting of the "Learning Forum" organized through the Prosperity Secretariat and the Conference Board of Canada in 1993, federal MP Bernard Valcourt stated that the legacy of the Prosperity Initiative was to " 'link competitiveness and learning in the minds of Canadians' " (Robertson, as cited in Christensen, 1998, p. 86). Christensen suggests that criticism of the CMEC's record by the federal Steering Group on Prosperity was one of the factors propelling the SAIP agenda forward.

The CMEC's emphasis on assessment-driven accountability derives largely from the current fixation with, and anxiety produced by, global market competition and the resultant need to know "exactly where one stands in comparison to the competition, both nationally and internationally, [as] the first step in becoming armed for the battle." (Christensen, 1998, p. 76) While the CMEC has been unresponsive to concerns about the SAIP from educational organizations including teachers' organizations, it has "chosen to privilege certain perspectives" (p. 92), namely those of business interests.

A Canadian Teachers' Federation (1992) draft position paper on accountability and testing in education raised some important issues with respect to how the various jurisdictions might respond to the initial results of the CMEC's national testing project and the longer-term, far-reaching impacts of the SAIP. It stated that:

> It is axiomatic that some provinces will rank lower than others (in fact, the results are almost predictable, on the basis of knowledge already available). What will these provinces do? Will they agree to adopt the curriculum and organizational strategies of the more successful provinces (thus moving towards a national core curriculum)? Will they attempt to modify existing curriculum so that it is more closely aligned with the tests? Will school boards seek to replicate the tests, thus making identification of individual schools and teachers more of a possibility? Will teachers be encouraged to teach to the next round of tests? Will the end result of the testing program be a pedestrian, overstandardized educational system in which there is no room for creativity or autonomy on the part of students or teachers and where the most neglected goals are those indeterminate aims of education which all agree are particularly important? (p. 44)

BEST COPY AVAILABLE

As the SAIP nears the end of its second full cycle and concerns begin to surface about aspects of the program (such as the process by which expectations for student performance are set), there are indications that it is influencing, as Robertson (1999) has observed, a "remarkable degree of curriculum homogenization within a country that has considered local control of curriculum to be a virtue rather than a liability." (p. 715) The creation by regional consortia in the Western and Atlantic provinces of outcomes-based curricula aligned with the SAIP reflects this trend.

The CTF paper goes on to emphasize the importance of having broad and diverse educational goals drive appropriate educational means:

> The fundamental error in the CMEC project is the failure to recognize that testing needs to be driven by the engines of educational goals and curriculum and instruction. Otherwise, the test results are likely to become an end in themselves and lead to undesirable practices in the schools .... while testing has a role to play in the development and improvement of education, it appears that the CMEC School Achievement Indicators Program falls into the category of testing which is mainly designed for monitoring and control. (pp. 44, 46)

Meaghan and Casas (1995b) describe the SAIP as a classic example of "placing the cart before the horse" – the development of a national testing program in the absence of public consensus on just what our educational goals should be and how these goals should be reflected in curriculum. This approach will result in "a handful of bureaucrats and consultants" deciding what it is they think students should know (p. 88).

As noted earlier, curriculum is becoming more test-driven and less responsive to student and community needs. Education means are dictating narrowly defined education goals. Barlow and Robertson note that standardized testing

BEST COPY AVAILABLE

> is a powerful force for curriculum reform and for shaping the goals of education. What has been and will be tested becomes a priority, not just in the public mind or ministry policy, but in time and attention in the classroom. Other instructional priorities recede, thus shifting curriculum balance effortlessly and without debate. (pp. 118-119)

This is being borne out south of the border. Over the past 150 years, testing requirements in the U.S. have had a number of unintended yet dramatic effects in terms of education policy. These include the reshaping of curriculum as well as attitudes about the goals of education (Madaus, 1994, p. 79). The impact of institutions such as the Educational Testing Service (ETS), the world's largest private testing firm, on education are significant and cannot be underestimated. Lemann (1995) wrote in the *Atlantic Monthly* that the

> ETS and its tests are a national obsession, deeply worked into the fabric of middle-class life. They have generated a large, independent test-prep industry and have strongly influenced elementary and secondary education, one of whose central goals is preparation for ETS tests. (p. 42)

The bigger picture of educational reform

The trends in assessment cited in the introduction to this report must be situated within the context of broader trends in educational change, among them: reduced funding, reforms increasingly being justified in economic terms (i.e. education as the means to prepare individuals for the labour market and nations for global competition), and a move toward market-based solutions which view education as a commodity and students and their parents as consumers of that commodity (Levin, 1998). Such trends, motivated more by fiscal, political and ideological considerations than by a sincere desire to improve access to high quality

education, are hardly surprising in a world in which the prevailing ideology is one of globalization.[11]

The influence of international bodies such as the Organisation for Economic Co-operation and Development (OECD) and the Asia-Pacific Economic Community (APEC) on education policy is being increasingly felt. In his analysis of the APEC position paper on education prepared by the South Korean government, Kuehn (1997b) notes that the "globalization agenda for education" relies on the premise that "globalization is inevitable and education [along with every other sector of society] must comply with its requirements". A major goal of education in the global economy is the preparation of workers for business, hence "business should have a central role in determining the content of schooling". The content of education becomes narrowed to work ethic, attitudes, and skills "appropriate for a working life". In this context standardized testing becomes a convenient means of regulating the production of the necessary "human capital". To illustrate, Kuehn (1997a) cites the case of Mexico where a standardized examination, the Examen Unico, has been created to channel more students into vocational/technical programs that produce the skills required for the global economy and away from university programs "primarily aimed at the specific social and cultural development of Mexico." (p. 86) Opposition to the Examen Unico was strong to say the least. In August 1996, anti-test demonstrators blocked busy streets in Mexico City and forced a one-day closure of the Mexican stock exchange.

Robertson's (1998a) analysis of high-stakes testing in the context of globalization illustrates well how testing is becoming an important lever for a number of major education reforms. Globalization is moving us inexorably towards what she describes as McWorld – a highly inequitable one-third/two-thirds world in which "[t]he one-

[11] For a summary of some of the main features of globalization, see "On Globalization" by Larry Kuehn, British Columbia Teachers' Federation [available online at: http://www.ctf-fce.ca/e/what/other/onglobe.htm].

third are those who will do reasonably well after international economic restructuring; the [other] two-thirds will not" and "everything is ordered to suit the market and the unbridled pursuit of profit." Noting that the "assault on public education is about creating McSchools to match and serve McWorld", she demonstrates how high-stakes testing can be used as the

> means to leverage support for any number of McSchool reforms that would otherwise be rejected as unconstitutional, ill-advised, or contrary to everything else we know about fostering good schools and good students. High-stakes testing is so closely and yet opaquely related to the other reforms of McSchools – the narrowing of curriculum, the deprofessionalization of teachers, privatization and compulsory enthusiasm for technology – that it has been strategically brilliant to focus on this strategy.

Mechanisms such as high-stakes standardized testing and performance indicators not only embody dominant cultural values associated with competition, efficiency, performance, and outcomes (the more concrete and measurable, the better). They are also an important link between the parallel and apparently contradictory tendencies in education reform of decentralization (weak state) and centralization (strong state). As decentralization takes shape in reforms such as site-based management in which decision-making responsibility is devolved to the school level, and school choice in which power is devolved to individual "consumers" of education, governments are centralizing power over funding, curriculum, standards and assessment. As noted earlier, assessment of centrally-set outcomes and the subsequent publication of results enables parents to make comparisons and select among schools – thus facilitating a measure of *accountability to the market* for the school system. In his review of Whitty, Power and Halpin's book *Devolution and Choice in Education: The School, the State and the Market*, Apple (1998) discusses the authors' concept of the "evaluative state":

> At the same time as the state appears to be devolving power to individuals and autonomous institutions that are themselves increasingly competing in a market, the state remains strong in key areas .... In many ways, this is actually one of the hallmarks of neo-liberalism. Indeed, one of the key differences between classical liberalism and its faith in 'enterprising individuals' in a market and current forms of neo-liberalism is the latter's commitment to a regulatory state. Neo-liberalism demands the constant production of evidence that one is in fact 'making an enterprise of oneself' .... Thus, under these conditions, not only does education become a market commodity like bread and cars in which the values, procedures, and metaphors of business dominate, but its results must be reducible to standardized 'performance indicators' (as quoted in Apple, 1998, p. 25).

In a market-based system, testing and performance indicators thus provide the necessary evidence that schools and school systems are "making an enterprise of themselves" thereby satisfying demands for market accountability – at the expense of genuine public accountability.

Some General Observations

Shifting from the global to the classroom context, it is important to emphasize that teachers recognize the importance of evaluating student learning and of finding the best means of classroom-based assessment to accomplish this. They regularly evaluate learner progress using various methods which they design and administer – these include term exams, quizzes, essays, projects, portfolios and classroom observation to name a few. Robertson (1998b) describes the diverse formal and informal evaluations carried out by teachers as being "remarkably good predictors of their students' future success, both in other courses and in life" (p. 65). According to Lorna Earl (1995), Professor of Theory and Policy Studies in Education at OISE/UT, "classroom teachers and their daily classroom assessments may form the basis for the most important accountability relationships." (p. 6) Because of their regular contact and close relationship with students, teachers are in the best position to conduct ongoing assessments of students, provide regular feedback on their performance to parents and to the students themselves, and recommend ways in which students can learn more effectively. Quoting from R. Stiggins' book *Student-Centered Classroom Assessment*, Earl (1995) makes the point that the rationale underlying the importance of classroom assessment in the improvement of learning is also the reason that large-scale assessments cannot achieve this:

> "...all of the centralized testing programs operate on the assumption that the decisions that exert the greatest influence on school quality are made at school management and policy levels; that is, somewhere outside the classroom ....This view of the origin of school improvement is naive and obviously counterproductive. The halls of political power are not where school

> improvement originates. There are three groups of decision-makers who determine the quality and therefore the impact of American schools. In order of importance, they are students themselves, their teachers and their parents. All three are informed in their decision-making by the same source of information on student achievement and that source is not national, state or local standardized testing. They all rely on teachers' day-to-day classroom assessments of student achievement. For this reason, the path to school improvement lies not through more standardized testing, but through the development and use of the best quality classroom assessment we can generate!" (as quoted in Earl, 1995, p. 7)

In addition to acknowledging the centrality of teachers to the assessment and evaluation process, we must recognize the value of the multiple goals of education – goals that go beyond simply preparing students to enter the labour force as important as that may be. There is also a need to evaluate evidence of *all* of these goals, not just cognitive abilities, with a view to improving the quality of education. Although effectively evaluating such goals as preparing students to assume the responsibilities of good citizenship is an important and complex task, "that complexity is precisely what makes those goals so important to teach." (Barlow & Robertson, p. 119)

Reliance on a single form of evaluation, particularly one as flawed as standardized tests, for making important decisions about students, teachers and schools is irresponsible. Equally important is the danger of relying on one form of evaluation for several different purposes (whether student progress, teacher performance, or program and system effectiveness). The correct evaluation tool needs to be used for the task at hand (Earl, 1995, p. 6).

We need to revisit the concept of accountability and what it actually means. Contrary to current trends in assessment policy, testing by itself does not constitute accountability. Darling-Hammond (1999) reminds us that accountability is

> a means by which individuals or organizations take responsibility for their actions so that those who depend on them can be assured that there are some safeguards in place to encourage good practices and to prevent bad practices or abuses, to have some course of redress for problems that arise, and to have some assurance of equitable and fair treatment.

With this in mind, she notes that in a genuine, comprehensive system of educational accountability, the various players, from the federal and provincial governments down to the individual school and teacher, have important roles and responsibilities they must fulfill and for which they must be held to account. For example, the state or province "should be responsible for resourcing schools so that they can provide adequate education, for ensuring that there is equity in the distribution of those resources, [and] for assuring that there is a means by which qualified, well-prepared staff will be in all schools in those settings." Local school districts are responsible "[f]or practices that are supportive of good teaching and learning in schools, [and] for processes that support [these practices]." Teachers are responsible for "making good educational decisions on behalf of children". Darling-Hammond (1999) describes the latter as professional accountability which means investing in teachers' skills and knowledge. This is a central component of a genuine accountability system, one which requires providing teachers with ongoing access to the knowledge and support necessary to make sound decisions in order to improve student learning. Unless and until we hold *all* of the actors accountable for their diverse responsibilities, we can only speak of accountability in a vacuum.

In sum, there is little evidence to support standardized testing from a learning standpoint. Moreover, standardized testing has been shown to have a negative impact on learning, and is particularly harmful to at-risk students – it perpetuates and worsens educational inequities. Assessment, including authentic assessment, must be viewed first and foremost as part of a diverse strategy to improve the quality of education and address equity issues – other systemic reforms must be

implemented in tandem that address broad social and educational conditions. If a primary goal of our education system is to offer all students equality of educational opportunity, then standardized testing will only move us further away from this goal.

Appendix A — Summary of Problems with Standardized Testing

A considerable body of research has identified numerous problems associated with the nature and use of standardized achievement tests. The following is a partial list (excerpted from Meaghan & Casas, "On the Testing of Standards and Standardized Achievement Testing: Panacea, Placebo, or Pandora's Box?" *Interchange*, 26(1), 1995, pp. 34-35):

- √ To permit machine scoring, standardized tests are limited to multiple-choice items which, in turn, severely limits what can be tested. For example, you can test language skills but not writing ability or how students use language.

- √ To ensure comparability of test scores over time, few substantive changes are made over the years, which hardly encourages updating of curricula.

- √ Although advocated as a means to carry out international comparisons of student learning and performance, these tests are not sensitive to differences in curricula for particular ages and grades in different countries (nor can they be made so).

- √ Because each question is assigned the same score value, the student who answers more of the difficult (and/or important from the program standpoint) questions is judged the equal of the student who gets only the easy ones.

- √ In order to achieve a proper dispersion in the scores [to facilitate making comparisons], questions that are almost always answered correctly or incorrectly tend to be avoided. Aspects of performance that should be tested may therefore not be represented at all because they make no contribution to score dispersion. This leads to overtesting of minutiae and undertesting of important conventions.

- √ Standardized tests stress the product of learning, not the process. They measure a student's ability to recall facts, define words, and perform routine calculations, not higher learning processes such as analyzing, synthesizing, forming hypotheses, and exploring alternative ways of solving problems.

- √ More crucially, standardized tests have nothing to do with performance standards but only measure relative standards.

- √ Standardized tests do not serve as a useful diagnostic tool because they do not show students where they went wrong; corrected tests are never returned, excluding the process of feedback and debate that is at the core of learning and education.

- √ Standardized tests systematically rule out students' independent creativity and criticism by eliminating the need to construct answers for themselves.

- √ Rather than increasing accountability, standardized tests merely shift it from teachers and school authorities to anonymous government officials or corporate bureaucrats who cannot be confronted or held accountable if tests are poorly constructed, administered, or marked.

- √ Many studies have shown that standardized tests are biased against socioeconomic, racial, and ethnic minority groups.

Appendix B — *Principles for Fair Student Assessment Practices for Education in Canada*

(Joint Advisory Committee)

In 1993 a Canadian working group convened by the Joint Advisory Committee of the University of Alberta's Centre for Research in Applied Measurement and Evaluation produced a document called *Principles for Fair Student Assessment Practices for Education in Canada*.[12] The principles and their related guidelines address both classroom assessments and large-scale standardized assessments developed outside the classroom (by commercial test publishers, ministries of education and school boards), and represent a broad consensus of the education community (participating organizations included the Canadian Education Association, Canadian School Boards Association and Canadian Teachers' Federation among others). Intended for both developers and users of assessments (the former being those "who construct assessment methods and people who set policies for particular assessment programs"; the latter are those "who select and administer assessment methods, commission assessment development services, or make decisions on the basis of assessment results and findings" (p. 3)), the principles include the following:

[12] In the U.S., a project by the National Forum on Assessment, a coalition of education and civil rights organizations, led to the development in 1995 of Principles and Indicators for Student Assessment Systems. The first principle asserts that the primary purpose of assessment is to improve student learning.

- √ Assessment methods should reflect the purpose and context of the assessment.

- √ Students should be given sufficient opportunity to demonstrate the knowledge, skills, attitudes, or behaviours being assessed.

- √ Procedures for judging or scoring student performance should be appropriate for the assessment method used, and should be consistently applied and monitored.

- √ Procedures for summarizing and interpreting assessment results should yield accurate and informative representations of a student's performance in relation to the goals and objectives of instruction for the reporting period.

- √ Assessment reports should be clear, accurate, and of practical value to the intended audience.

Appendix C — CTF Principles on Curriculum, Assessment and Evaluation

The Canadian Teachers' Federation has developed its own set of principles on assessment and evaluation. In early 1998 the CTF held a workshop on curriculum and evaluation (referred to in the introduction) for the leadership of teacher organizations across the country. An important outcome of the workshop was the creation of draft statements of principle on assessment and evaluation as well as on curriculum. These statements of principle are intended to guide the work of provincial and territorial teacher organizations in a number of areas including the development, implementation and scope of curriculum, and high-stakes testing and its use. They have since been adopted as CTF policy and are reproduced here:

Principles on Curriculum

- √ Curriculum must be broad in scope and holistic in nature in order to meet the needs of children, community and the larger society.
- √ Curriculum must reflect realistic expectations based on knowledge of child development.
- √ Curriculum must promote, support, enhance and model equity and respect for diversity.
- √ Curriculum must be designed to prepare students to become caring, responsible and participatory members of society.

- √ Curriculum is ongoing and dynamic.
- √ Curriculum development, implementation and evaluation are multi-phased processes that require extensive time and resources.
- √ Teachers must have the primary role in, and control of, curriculum development.
- √ Professional dialogue and collaborative activity are essential for professional growth and development.
- √ Teachers require time and resources to engage in ongoing professional development, and in order to effectively implement curriculum change.
- √ Teachers must have the autonomy to choose the methods, timing and curriculum adjustments that are necessary to adapt the curriculum in order to meet students' needs.
- √ The goals society sets for students and schools must be challenging but attainable, and progress towards these goals must be measured fairly and in context.
- √ The conditions under which teaching and learning take place have a direct effect on what teachers and students can achieve together.
- √ The process for curriculum development must include opportunity for input from the broadest possible community representation.

Principles on Assessment and Evaluation

- √ Assessment is the gathering of reliable information pertaining to students' knowledge and understanding including social, economic and educational factors and resources that influence student and system performance.

- √ Students should be engaged in self- and peer-assessment, and should be fully informed of the purpose of assessments and evaluation.

- √ The assessment and evaluation of students, whether carried out at the classroom level or beyond, must take account of gender differences as well as cultural, linguistic, socio-economic and other forms of diversity.

- √ There are many forms, methods and approaches to assessment, none of which, in isolation, is an adequate measure of student, program or system performance.

- √ Information derived from assessment and evaluation of students should be collected, stored, reported and utilized in a manner respectful of students' rights to protection of privacy.

- √ Evaluation is the process of making judgements, based in part on assessment data.
- √ Assessment and evaluation are continuous processes that inform decision-making at classroom and system levels.
- √ Classroom teachers have the primary responsibility for assessing and evaluating student achievement.
- √ Governments and school systems have a collective responsibility to provide in-service opportunities for teachers in the areas of assessment and evaluation.
- √ Teachers' organizations, as the professional voice of teachers, must be directly involved in all policy decisions about assessment and evaluation at provincial/territorial, national and international levels.
- √ The purpose of student assessment is to support student learning.
- √ The design, interpretation and application of information derived from assessment and evaluation are complex tasks linked to professional beliefs, skills and knowledge.
- √ Decisions regarding evaluation and assessment should focus on what is valuable and not simply on what is measurable.
- √ The purposes and methods of assessing and evaluating student learning, curriculum, programs and systems are distinct and not interchangeable.
- √ The tools and processes used to assess and evaluate student learning must be congruent with the learning theory that is reflected in the curriculum.

References

Alberta Education. (1999, Mar. 11). "Details of $600 Million Boost for K-12 Education Announced." [News release] Edmonton. Available: http://ednet.edc.gov.ab.ca/news/1999nr/Mar99/nr-budgetpr.htm

"Alberta Proposes Bonuses for Student Achievement Gains." *Canadian Principal*, 10(7), April 1999, pp. 1-4.

American Association of University Women Educational Foundation. (1998). *Gender Gaps: Where Schools Still Fail Our Children*. Washington, DC: Author.

Apple, M. (1998, Aug./Sept.). "Are Markets and Standards Democratic?" [Book review.] *Educational Researcher*, 27(6), pp. 24-28.

Barker, K. C. (1991, Dec. 10). "Accountability as an Educational Reform Strategy: The Case of the School Achievement Indicators Program." Unpublished paper, Department of Educational Administration, University of Alberta, Edmonton.

Barlow, M., & Robertson, H.-j. (1994). *Class Warfare: The Assault on Canada's Schools*. Toronto, ON: Key Porter Books.

Berthelot, J. (1995, May). *On a Democratic Course or Drifting to Market Models?* Paper prepared for the Canadian Teachers' Federation National Conference "Public Education: Meeting the Challenges", Montreal, Quebec, May 1995.

Canadian Teachers' Federation. (1997, Sept.). *Behind the Charter School Myths*. Research Paper prepared for the National Issues in Education Initiative. Ottawa, ON: Author.

Canadian Teachers' Federation. (1995a, Oct.). "Canada's School Performance: Do International Comparisons Make Sense?" (Issue Sheet No. 9 – National Issues in Education Initiative).

Canadian Teachers' Federation. (1995b, Oct.). "National Testing." (Issue Sheet No. 1 – National Issues in Education Initiative).

Canadian Teachers' Federation. (1992, Feb.). "Accountability and Testing in Education, Draft Position Paper."

Casas, F., & Meaghan, D. (1995, March/April). "Standardized Tests: Why They Cannot Measure Educational Quality." *FWTAO Newsletter*, 13(4), pp. 16-21.

Christensen, L. (1998, Feb. 3). "Governing Under an Exclusive Lock and Key: The Council of Ministers of Education, Canada, 1967-1997." Unpublished paper, Faculty of Graduate Studies and Research (Canadian Studies), Carleton University, Ottawa.

Darling-Hammond, L. (1999, March). "Developing a Professional Model of Accountability for Our Schools." [online] Speech presented at the B.C. Teachers' Federation AGM, March 1999. Available: http://www.bctf.bc.ca/parents/speeches/darling-hammond.html

Darling-Hammond, L. (1994, Spring). "Performance-Based Assessment and Educational Equity." *Harvard Educational Review*, 64(1), pp. 5-30.

Darling-Hammond, L. (1991, Nov.). "The Implications of Testing Policy for Quality and Equality." *Phi Delta Kappan*, 73(3), pp. 220-225.

Dehli, K. (1998). "Shopping for Schools: The Future of Education in Ontario?" *Orbit*, 29(1), pp. 29-33.

Dunning, P. (1997, July). *Education in Canada: An Overview*. Toronto, ON: Canadian Education Association.

Earl, L. M. (1995, May). *Accountability in Education: Where Do Teachers Fit?* Paper prepared for the Canadian Teachers' Federation National Conference "Public Education: Meeting the Challenges", Montreal, Quebec, May 1995.

Education Week on the Web (in collaboration with the Pew Charitable Trusts). (1999, Jan. 11). *Quality Counts '99: Rewarding Results, Punishing Failure*. Editorial Projects in Education. [online] Available: http://www.edweek.org/sreports/

Eisner, E. (1998, March). "The Role of Teachers in Assessment and Education Reform." [online] Speech presented at the B.C. Teachers' Federation AGM, March 1998. Available: http://www.bctf.bc.ca/parents/speeches/eisner.html

FairTest (National Center for Fair & Open Testing). (1995). *Selected Annotated Bibliography on the SAT (Scholastic Aptitude Test): Bias and Misuse*. Cambridge, MA: Author.

Gardner, H. (1999, Feb.). "Who Owns Intelligence?" *Atlantic Monthly*, 283(2), pp. 67-76.

Gilliss, G. (1994). *Province-Wide Assessment Programs*. Ottawa, ON: Canadian Teachers' Federation.

Gilliss, G. (1993, Jan. 12). "Why National Tests Won't Improve Education." *Globe and Mail*.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991, June/July). "Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution." *Educational Researcher*, 20(5), pp. 2-7.

Hart, D., & Livingstone, D. W. (1998, Spring). "The 'Crisis' of Confidence in Schools and the Neoconservative Agenda: Diverging Opinions of Corporate Executives and the General Public." *The Alberta Journal of Educational Research*, XLIV(1), pp. 1-19.

"Improvement, Not Incentives: Joint Proposal to Replace School Performance Incentive Program with Alberta School Improvement Program." [online] Jointly submitted to the Minister of Education by the Alberta Home and School Councils' Association, The Alberta School Boards Association, the Alberta Teachers' Association and the College of Alberta School Superintendents, May 1999. Available: http://www.teachers.ab.ca/what/asip.html

Ireland, D. S. (1997a, June 4). *Analysis of the Tasks Used in the Grade 3 Provincial Assessment of Mathematics and Language Arts 1997*. Prepared for the Ontario English Catholic Teachers Association.

Ireland, D. S. (1997b, Spring). "Canadian Students Do Better Than Results Indicate." *CCPA Education Monitor*, 1(2), pp. 13-16.

Ireland, D. S. (1995, May). *The International Status of Canadian Education* (Revised). Carleton Board of Education.

Jones, K., & Whitford, B. L. (1997, Dec.). "Kentucky's Conflicting Reform Principles: High-Stakes School Accountability and Student Performance Assessment." *Phi Delta Kappan*, 70(4), pp. 276-281.

Kleinsasser, A. (1994, Spring). "Book Review – *Testing Testing: Social Consequences of the Examined Life* by F. Allan Hanson." *Harvard Educational Review*, 64(1), pp. 96-102.

Kuehn, L. (1997a, April/May). "Parents, Standardized Testing and the Mexican Stock Exchange." *Our Schools/Our Selves*, 8(3), #51, pp. 83-90.

Kuehn, L. (1997b). "Schools for Globalized Business: The APEC Agenda for Education – A Commentary on the 'Concept Paper' for the APEC Human Resources Ministerial Meeting." [online] Discussion paper prepared for the Public Education and Research Forum, 1997 People's Summit on APEC, Vancouver, BC, Nov. 19-20, 1997. Available: http://www.vcn.bc.ca/ed8resea/larry.html

Lemann, N. (1995, Aug.). "The Structure of Success in America." *Atlantic Monthly*, 276(2), pp. 41-60.

Levin, B. (1998, June). "An Epidemic of Education Policy: (What) Can We Learn From Each Other?" *Comparative Education*, 34(2), pp. 131-141.

Madaus, G. F. (1994, Spring). "A Technological and Historical Consideration of Equity Issues Associated with Proposals to Change the Nation's Testing Policy." *Harvard Educational Review*, 64(1), pp. 76-95.

Madaus, G. F. (1991, Nov.). "The Effects of Important Tests on Students: Implications for a National Examination System." *Phi Delta Kappan*, 73(3), pp. 226-231.

Madaus, G. F., & Kellaghan, T. (1993, April). "Testing as a Mechanism of Public Policy: A Brief History and Description." *Measurement and Evaluation in Counseling and Development*, 26(1), pp. 6-10.

McLaughlin, M. W. (1991, Nov.). "Test-Based Accountability As a Reform Strategy." *Phi Delta Kappan*, 73(3), pp. 248-251.

McMahen, L. (1999, March 15). "Alberta to Grant Bonus Money to Boards for Performance." *The Edvisor*, 1(23).

McMurtry, J. (1992, Spring). "National Testing." *Canadian Social Studies*, 26(3), pp. 94-95.

Meaghan, D. E., & Casas, F. R. (1995a, Oct./Nov.). "Let's Look Before We Leap: Standardized Testing and Special Education Students." *CEA Newsletter*, p. 5.

Meaghan, D. E., & Casas, F. R. (1995b). "On Standardized Achievement Testing: Response to Freedman and Wilson and a Last Word." *Interchange*, 26(1), pp. 81-96.

Meaghan, D. E., & Casas, F. R. (1995c). "On the Testing of Standards and Standardized Achievement Testing: Panacea, Placebo, or Pandora's Box?" *Interchange*, 26(1), pp. 33-58.

Meaghan, D. E., & Casas, F. R. (1995d, Sept.). "Quality Education and Other Myths: A New Face for an Old Conservative Agenda." *Our Schools/Our Selves*, 7(1), #43, pp. 37-53.

Medina, N., & Neill, D. M. (1990, March). *Fallout From the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools* (Third Edition - Revised). Cambridge, MA: National Center for Fair & Open Testing (FairTest).

Nagy, P. (1996, Feb.). *International Comparisons of Student Achievement in Mathematics and Science*. Ontario Institute for Studies in Education.

Nagy, P. (1995, June). "International Comparisons of Student Achievement: Examining the Tests and the Data" (Draft). Presented at the Annual Conference of the Canadian Society for the Study of Education, Montreal, June 1995.

National Advisory Board on Science and Technology. (1994, May). *National Standards in Education: A Question of Excellence. Report of the National Advisory Board on Science and Technology Committee on National Standards in Education*. Ottawa.

National Forum on Assessment. (1995). *Principles and Indicators for Student Assessment Systems*. Cambridge, MA: National Center for Fair & Open Testing (FairTest).

Neill, M., et al. (1997, Sept.). *Testing Our Children: A Report Card on State Assessment Systems*. Cambridge, MA: National Center for Fair & Open Testing (FairTest).

Oakes, J. (1985). *Keeping Track: How Schools Structure Inequality*. New Haven, CT: Yale University Press.

Peirce, B. N., & Stein, P. (1995, Spring). "Why the 'Monkeys Passage' Bombed: Tests, Genres, and Teaching." *Harvard Educational Review*, 65(1), pp. 50-65.

Principles for Fair Student Assessment Practices for Education in Canada. (1993). Edmonton, AB: Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation, University of Alberta.

Reardon, S. F., Scott, K., & Verre, J. (1994, Spring). "Symposium: Equity in Educational Assessment – Introduction." *Harvard Educational Review*, 64(1), pp. 1-4.

Rehnby, N. (1996, Feb.). *Inventing Crisis: The Erosion of Confidence in Canadian Public Education*. Prepared for the Canadian Teachers' Federation by the British Columbia Teachers' Federation.

Robertson, H.-j. (1999, May). "Bogus Points." ["In Canada" column.] *Phi Delta Kappan*, 80(9), pp. 715-716.

Robertson, H.-j. (1998a). "McSchool Meets McWorld: The Role of High-Stakes Testing." Paper prepared for the Manitoba Teachers' Society, March 5, 1998.

Robertson, H.-j. (1998b). *No More Teachers, No More Books: The Commercialization of Canada's Schools*. Toronto, ON: McClelland & Stewart.

Robertson, H.-j. (1992, Nov.). *The Better Idea Book: A Resource Book on Gender, Culture, Science and Schools*. Ottawa, ON: Canadian Teachers' Federation.

Rosser, P., et al. (1992, Aug.). *Sex Bias in College Admissions Tests: Why Women Lose Out* (4th Edition). Cambridge, MA: National Center for Fair & Open Testing (FairTest).

Schwartz, T. (1999, Jan. 10). "The Test Under Stress." *New York Times Magazine*.

Thurmond, M. (1994, Spring). "The FairTest Case: Nothing Special, Just Equality." *Civil Liberties* (National Newsletter of the American Civil Liberties Union), #380.

Torrance, H. (1993, Spring). "Combining Measurement-Driven Instruction with Authentic Assessment: Some Initial Observations of National Assessment in England and Wales." *Educational Evaluation and Policy Analysis*, 15(1), pp. 81-90.

Traub, R. (1994). *Standardized Testing in Canada: A Survey of Standardized Achievement Testing by Ministries of Education and School Boards*. Toronto, ON: Canadian Education Association.

Weiss, J. G. (1991). "It's Time to Examine the Examiners." In A. G. Hilliard III (Ed.), *Testing African American Students – Special Re-Issue of The Negro Educational Review* (pp. 107-124). Morristown, NJ: Aaron Press.

Wideen, M. F., O'Shea, T., Pye, I., & Ivany, G. (1997). "High-Stakes Testing and the Teaching of Science." *Canadian Journal of Education*, 22(4), pp. 428-444.

LES TESTS STANDARDISÉS

ATTEINTE À L'ÉQUITÉ EN ÉDUCATION

Bernie Froese-Germain

BEST COPY AVAILABLE

72

dération canadienne des enseignantes et des enseignants

Les tests standardisés : atteinte à l'équité en éducation

Rapport préparé dans le cadre de la campagne des Dossiers nationaux en éducation

par Bernie Froese-Germain
Fédération canadienne des enseignantes et des enseignants

Novembre 1999

73

Pour obtenir plus de renseignements, veuillez vous adresser à la :

Fédération canadienne des enseignantes et des enseignants
110, avenue Argyle
Ottawa ON K2P 1B4
Tél. : (613) 232-1505
Téléc. : (613) 232-1886
Courriel : info@ctf-fce.ca
Internet : http://www.ctf-fce.ca

La Fédération canadienne des enseignantes et des enseignants a été créée en 1920 pour veiller à ce que les autorités extraprovinciales tiennent compte de l'opinion des membres de la profession enseignante lorsqu'elles envisagent des mesures touchant la profession ou son travail auprès des élèves. Porte-parole nationale de la profession enseignante, la FCE fait la promotion, aux échelons interprovincial, national et international, de la qualité de l'éducation, de la situation des membres de la profession et de l'égalité des chances au moyen de l'éducation. La FCE coordonne et facilite la mise en commun d'idées, de connaissances et de compétences parmi les organisations provinciales et territoriales qui lui sont affiliées et qui regroupent au total 240 000 enseignantes et enseignants.

ISBN: 0-88989-326-8

© *Fédération canadienne des enseignantes et des enseignants, 1999*

Cette publication, y compris la traduction, la mise en page et l'impression, est le fruit de la collaboration du personnel de la FCE.
Couverture conçue par Prosar Canada Inc.

Table des matières

Avant-propos i

Introduction 1
Les problèmes que posent les tests standardisés — Aperçu 5
Incapacité d'évaluer l'apprentissage et le développement de l'élève . 7
Autres problèmes 15

Les tests et l'iniquité en éducation 17
Mauvaise utilisation des tests 17
Responsabilité à grands enjeux axée sur la mesure 23
La partialité des tests 30

Quelques mots sur l'évaluation authentique 39

La politique des tests 43
Comparaisons internationales 44
Les tests, les écoles et la concurrence du marché 47
La « pancanadianisation » du curriculum et de l'évaluation 49
Une vision d'ensemble de la réforme de l'éducation 55

Observations générales 61

Annexe A –
Sommaire des problèmes que posent les tests standardisés 65
Annexe B –
Principes d'équité relatifs aux pratiques d'évaluation des apprentissages scolaires au Canada (Comité consultatif mixte) 69
Annexe C –
Principes de la FCE sur le curriculum, la mesure et l'évaluation ... 71
Principes sur le curriculum 71
Principes sur la mesure et l'évaluation 73

Références 75

Avant-propos

Les tests de rendement des élèves foisonnent dans les écoles canadiennes. Par suite des changements apportés aux politiques d'évaluation, le recours aux tests standardisés aux échelons provincial, national et international va grandissant, des élèves d'un plus vaste éventail d'années d'études et en plus bas âge sont soumis à des tests, les résultats des tests sont diffusés plus largement et on y attache de plus en plus d'importance. Les adeptes prétendent qu'en administrant plus de tests, on accroîtra la responsabilisation des écoles — le testage semble s'imposer comme synonyme de responsabilisation.

Devant cette mise en valeur du testage comme instrument de réforme éducationnelle, dans sa publication intitulée *Les tests standardisés : atteinte à l'équité en éducation*, la FCE soutient que l'emploi de tests standardisés se justifie mal du point de vue pédagogique et étudie les effets qu'ont les tests sur l'équité en éducation. Les problèmes que posent les tests standardisés sont bien documentés. Ils tiennent autant à la nature des tests qu'à leur utilisation et à l'interprétation de leurs résultats. Dans son rapport, la FCE analyse certains de ces problèmes, en s'attachant particulièrement aux répercussions sur l'équité en éducation de la partialité des tests et du mauvais emploi de leurs résultats. Elle examine également la mesure axée sur le rendement, de même que la politique entourant le testage standardisé, formule des observations générales sur la mesure de l'apprentissage des élèves et fournit des principes établis par des éducateurs et éducatrices pour guider la prise de décisions en matière de mesure, d'évaluation et de curriculum.

Destinée à faire prendre conscience des questions cruciales liées aux effets du testage standardisé, cette publication constituera une

ressource utile pour les enseignantes et enseignants à tous les paliers, ainsi que pour les administrateurs et administratrices, les recherchistes, les décisionnaires et quiconque s'intéresse à la mesure et à l'évaluation du rendement des élèves.

Bernie Froese-Germain est recherchiste à la Fédération canadienne des enseignantes et des enseignants à Ottawa.

Introduction

Dans une conférence prononcée lors d'un atelier sur le curriculum et l'évaluation organisé par la Fédération canadienne des enseignantes et des enseignants en 1998, M. Benjamin Levin, doyen de l'éducation permanente à l'University of Manitoba, a donné un aperçu des changements qui s'opèrent dans tout le pays au chapitre des politiques en matière d'éducation.[1] En ce qui concerne le curriculum, les politiques sont davantage axées sur les « matières de base » et elles prévoient moins d'options, une répartition plus restrictive du temps, l'élaboration de programmes régionaux et pancanadiens, ainsi que des périodes de mise en œuvre plus courtes pour les nouveaux curriculums. En ce qui concerne la mesure, les politiques préconisent l'emploi accru du testage aux échelons provincial, national et international, une diffusion plus générale des résultats des tests, malgré la piètre qualité des rapports livrés au public, et une importance plus grande donnée à ces résultats. Le Canada n'est d'ailleurs pas le seul pays à accorder plus d'importance au testage. Selon Levin (1998),

> les normes, la responsabilisation et le testage sont désormais au centre des réformes mises en œuvre par de nombreux pays. Presque partout, on constate depuis

[1] Le colloque « Programme d'études et évaluation : Organisations de l'enseignement et nouveautés pancanadiennes » a été organisé par la Fédération canadienne des enseignantes et des enseignants en janvier 1998, dans le cadre de la campagne des Dossiers nationaux en éducation. Il avait pour but , de réunir des dirigeants d'organisations de l'enseignement afin de discuter des dernières innovations en matière de curriculum et d'évaluation, et particulièrement de leur nature de plus en plus « pancanadienne ». Les participants et participantes ont élaboré des déclarations de principe sur le curriculum, la mesure et l'évaluation, qui sont reproduites à l'annexe C du présent rapport.

> quelques années une utilisation accrue du testage à grande échelle et une diffusion plus large des résultats de ces tests. Cette tendance à l'échelon national s'accompagne d'une augmentation des mesures effectuées à l'échelon international; dans un cas comme dans l'autre, les résultats servent de plus en plus ouvertement à établir des comparaisons publiques★ (p. 133).

Les tenants et tenantes des tests standardisés prétendent que ceux-ci permettent de responsabiliser davantage le personnel enseignant et les écoles. Non seulement ces affirmations sont peu convaincantes, mais nombreuses sont les données montrant les multiples problèmes liés à l'emploi des tests standardisés (Meaghan et Casas, 1995d, p. 46). Ces problèmes sont attribuables non pas tant à la nature (forme et contenu) des tests qu'à leur utilisation et à l'interprétation de leurs résultats. Le présent rapport analyse certains de ces problèmes, en s'attachant particulièrement aux conséquences de la partialité des tests et du mauvais emploi de leurs résultats sur l'équité en éducation. Il traite aussi brièvement de la mesure authentique ou mesure du rendement, étudie le testage standardisé dans une perspective politique et formule des observations générales sur la mesure de l'apprentissage et du développement des élèves.[2] Les annexes jointes au rapport renferment des principes élaborés par des éducateurs et éducatrices pour guider la prise de décisions en matière de mesure, d'évaluation et de curriculum.

---

[2] Même si les termes *mesure* et *évaluation* sont souvent utilisés de façon interchangeable (notamment dans le présent rapport), il existe une distinction subtile mais néanmoins importante entre ces deux notions. On entend par mesure la collecte de renseignements fiables concernant la compréhension et les connaissances acquises par les élèves, y compris les ressources et les facteurs sociaux, économiques et éducationnels qui influent sur le rendement des élèves et du système. L'évaluation désigne la formulation de jugements fondés en partie sur les données établies par la mesure.

★ Partout dans le texte, l'astérisque sert à indiquer qu'il s'agit d'une traduction de la FCE.

Les nombreuses allusions du rapport aux tests employés aux États-Unis reflètent l'expérience étatsunienne considérable dans ce domaine, expérience dont le Canada peut tirer d'importantes leçons. On estime que les écoles publiques aux États-Unis ont administré plus de 200 millions de tests standardisés au cours de l'année scolaire 1986-1987 (Medina et Neill, 1990). Selon Barlow et Robertson (1994), « si les tests étaient synonymes d'excellence, les élèves des États-Unis, les "plus testés au monde", seraient les grands gagnants de la loterie de l'éducation »★ (p. 117); en réalité, les États-Unis ont régulièrement des résultats médiocres dans les évaluations internationales à grande échelle (Earl, 1995, p. 5). Les recherches effectuées au cours des deux dernières décennies confirment que les tests standardisés ont entraîné « de nombreuses conséquences négatives sur la qualité de l'éducation aux États-Unis et sur l'équité des chances en éducation »★
(Darling-Hammond, 1991, p. 220).

Les problèmes que posent les tests standardisés — Aperçu[3]

Mis au point pendant la Première Guerre mondiale afin de repérer les recrues militaires ayant le potentiel de devenir officiers, les tests standardisés servaient à répartir les candidats afin de déceler les plus aptes à assumer un rôle de leader (Casas et Meaghan, 1995, p. 18).[4] Voici la définition des tests standardisés de rendement que donne Traub (1994) :

> De façon générique, un test de rendement a pour but d'évaluer les connaissances et la compréhension qu'un élève a d'un sujet scolaire donné. Un test normalisé de rendement se définit, de surcroît, comme étant administré et évalué de la même manière quels que soient son lieu et sa date d'utilisation. La normalisation implique que les résultats de tous les élèves peuvent être comparés les uns avec les autres de façon impartiale. [...] l'essentiel est que les conditions d'administration et d'évaluation soient les mêmes pour tous les élèves de façon que leurs résultats puissent être comparés (p. 5-6).

[3] L'Annexe A résume certains des principaux problèmes que posent les tests standardisés.

[4] Le Scholastic Aptitude Test (SAT) employé aux États-Unis a été mis au point dans les années 20 à partir du test « Army Alpha » qui servait à sélectionner d'éventuels officiers parmi les soldats (Lemann, 1995). Madaus et Kellaghan (1993) présentent un bref historique de l'administration des tests en général (y compris l'introduction du classement au XVIII[e] siècle et l'invention des notes chiffrées).

Ils sont fréquemment appelés « tests normatifs » parce que les résultats d'un élève sont comparés à ceux d'un groupe représentatif plus important (qui constitue la « norme »). Le test se présente souvent sous forme de questions à choix multiple, ce qui se prête très bien à la correction assistée par ordinateur. Mis à part le perfectionnement des méthodes électroniques de notation, d'analyse et de communication des résultats, dans l'ensemble les tests standardisés de rendement ont très peu changé au cours des 50 dernières années, même si l'on dit qu'ils sont maintenant plus « scientifiques » (Medina et Neill, 1990, p. 13) et même si nous savons mieux en quoi consiste un enseignement et un apprentissage efficaces.

Une abondante recherche documente les multiples défauts que présentent l'élaboration, la validation, l'administration et l'utilisation des tests standardisés. Selon Medina et Neill, de FairTest (National Center for Fair & Open Testing in Massachusetts) :

> La structure des tests standardisés garantit souvent des résultats faussés à l'égard des minorités, des filles et des élèves à faible revenu. Les résultats sont évalués et cotés d'une façon qui n'est pas conforme aux théories modernes sur l'intelligence et le développement de l'enfant. Le processus de validation des tests est souvent inadéquat et très peu objectif. Dans de nombreux cas, les tests sont administrés dans des conditions qui sont loin de satisfaire aux exigences de la « standardisation ». Même s'ils sont administrés de façon « standard », leurs résultats peuvent encore être faussés à l'endroit des minorités, des élèves à faible revenu et des filles, en ayant recours à des examinateurs inconnus des élèves et à des tests pour lesquels le temps alloué est très serré. En dépit de ce que prétendent les concepteurs de ces tests, ces défauts rendent l'objectivité impossible et font que les

résultats des tests sont inexacts, peu fiables ou faussés. En fin de compte, de nombreux tests ne mesurent pas efficacement le rendement, les aptitudes ou les compétences des personnes qui y sont soumises* (p. 8).

Incapacité d'évaluer l'apprentissage et le développement de l'élève

Les tests standardisés sont peut-être utiles au classement des élèves, mais ils n'évaluent pas adéquatement leur apprentissage et leur développement. Selon la documentation qui existe sur le sujet, cette faiblesse est attribuable aux causes suivantes :

De nombreuses aptitudes ne peuvent être évaluées par un test standardisé.

Des qualités comme le sens civique, l'éthique, la confiance et l'estime de soi, l'appréciation esthétique, le respect des autres, la discipline, l'aptitude sociale et le désir d'apprendre contribuent énormément à l'atteinte des buts multiples de l'éducation, même si elles ne sont pas facilement quantifiables. En obligeant les écoles et le personnel enseignant à se concentrer sur des compétences fondamentales et quantifiables, le testage standardisé constitue une sérieuse menace pour la diversité des buts et du contenu de l'éducation. Medina et Neill nous disent que « les tests standardisés ne peuvent évaluer que de façon limitée la connaissance, les aptitudes et les compétences des élèves, car ils sont fondés sur l'hypothèse que l'on peut isoler ces attributs les uns des autres, les ordonner sur une échelle linéaire et les exprimer sous forme d'une seule note »* (p. 12).

Dans son ouvrage précurseur *The Mismeasure of Man*, Gould dément deux grands mythes sur l'intelligence humaine (cité dans Darling-Hammond, 1994, p. 10-11) : l'intelligence est un « produit inné, simple et mesurable » et, en partant de la notion que l'intelligence pourrait être quantifiée de cette manière, les personnes peuvent être ainsi comparées et classées. Différentes théories ont été élaborées

sur la notion d'intelligence. Gardner (1999) note qu'on discute, entre autres, de « ce qu'est l'intelligence, comment (et si) elle doit être mesurée et en fonction de quelles valeurs l'intellect humain doit être mesuré » (p. 70). D'après lui, tout le monde « possède au moins huit formes d'intelligence : linguistique et logico-mathématique (les deux formes les plus valorisées à l'école et celles qui sont essentielles à la réussite des tests standardisés d'intelligence), musicale, spatiale, corporelle-kynesthésique, naturaliste, interpersonnelle et intrapersonnelle »; et « personne ne possède les mêmes formes d'intelligence au même degré » (p. 71). Ces théories ont naturellement des répercussions considérables sur les méthodes de mesure et d'évaluation scolaires.

Les tests ont beau être standardisés, les élèves ne le sont pas.

La capacité d'apprentissage d'un élève est influencée par toute une gamme de facteurs, notamment : l'incidence de la pauvreté; le niveau de scolarité des parents; la santé mentale, physique et affective; les effets du racisme et d'autres formes de discrimination, et la langue d'origine. Par conséquent, l'absence de renseignements sur les caractéristiques des élèves, dans les écoles qui participent aux programmes de tests, mène à une utilisation trompeuse et peu judicieuse des données recueillies (Fédération canadienne des enseignantes et des enseignants, 1995b). Lors d'une enquête effectuée en 1994 par la Fédération canadienne des enseignantes et des enseignants sur les programmes d'évaluation administrés à l'échelle provinciale, quelques administrations ont dit recueillir certains renseignements généraux sur les élèves (principalement d'ordre démographique) dans le cadre des tests. Dans une province, les renseignements obtenus portaient sur des facteurs comme la participation des parents aux études de l'élève et la quantité de devoirs. Cependant, aucune administration n'a recueilli de renseignements sur l'origine socioéconomique des élèves (Gilliss, 1994, p. 5).

Selon Gilliss (1993), « ceux qui s'attendent à ce que les élèves désavantagés, pauvres, mal nourris, négligés ou gravement handicapés obtiennent les mêmes résultats que les élèves qui vivent dans l'affluence, sont en bonne santé, sont bien choyés et n'ont aucune

déficience [physique] ne peuvent qu'être déçus. »* Elle fait également remarquer que la moyenne des résultats des tests administrés dans les écoles contient plus de renseignements sur les caractéristiques regroupées des élèves que sur les écoles. Au lieu d'engager des dépenses pour concevoir, administrer et noter les tests afin de savoir quelles écoles obtiennent les meilleures notes, on pourrait obtenir des renseignements très semblables en faisant l'analyse socioéconomique des données de recensement sur les collectivités où se trouvent ces écoles. McLaughlin (1991) est du même avis. En réponse à la demande de tests standardisés dans « America 2000 » (la stratégie nationale des États-Unis pour l'éducation, annoncée sous le mandat du président George Bush), elle fait remarquer que le testage

> exige que l'on rende compte des extrants sans égard à la responsabilité de la société en ce qui concerne les intrants. Par conséquent, même les meilleures stratégies en matière de testage ne font que démontrer les inégalités considérables qui existent dans tout le système scolaire étatsunien et ne prévoient rien pour combler les besoins fondamentaux en ressources suffisantes à l'appui de ce travail* (p. 250).

Les tests standardisés conçus pour être administrés à un grand nombre d'élèves doivent nécessairement être de nature très générale.

Il en résulte un écart entre la matière enseignée (compte tenu de la diversité des curriculums) et la matière évaluée par les tests (Casas et Meaghan, 1995, p. 18-19). Ireland (1997b) analyse les résultats de 8e année recueillis dans le cadre de la Troisième étude internationale des mathématiques et des sciences (TEIMS) publiée en novembre 1996 et souligne les différences importantes qui existent entre les provinces en ce qui concerne la concordance entre les tests et les curriculums. Par exemple, les conclusions de la TEIMS indiquent qu'en Colombie-Britannique, 98 p. 100 des questions de tests correspondent au curriculum des sciences, tandis qu'en Ontario, en Alberta et à Terre-Neuve, cette proportion n'est que de

53 p. 100, 54 p. 100 et 54 p. 100 respectivement. Cependant, l'étude donne peu d'explications sur cette concordance entre les tests et le curriculum en dehors des tableaux et des chiffres qu'elle renferme. Selon Ireland, l'étude aurait été plus pertinente si on avait posé les questions suivantes (Ireland, 1997b, p. 14) : Dans quelle mesure les élèves ont-ils su répondre aux questions portant sur le curriculum qui était censé leur être enseigné? Dans quelle mesure le test était-il représentatif du curriculum de chaque administration et quelle était l'importance du volet du curriculum qui n'a pas été évalué par le test? Quels résultats évalués par le test ne faisaient pas partie du curriculum de chaque administration et quelle était leur importance?

Les tests standardisés mesurent normalement le rappel élémentaire de faits et de compétences et pénalisent la réflexion d'ordre supérieur.

Selon Meaghan et Casas (1995d), les tests standardisés « mesurent l'aptitude de l'élève à se rappeler des faits, à définir des mots et à accomplir des activités routinières, plutôt que des compétences d'ordre supérieur, comme l'analyse, la synthèse, la formulation d'hypothèses et l'exploration de nouvelles façons de résoudre des problèmes ». De plus, ils « excluent le processus de rétroaction et de discrimination qui est au cœur même de l'apprentissage [...et] rejettent d'emblée la créativité et la critique indépendantes en abolissant complètement le besoin de structurer les réponses »* (p. 47). McMurtry (1992) illustre cette faiblesse en analysant une question tirée d'un test de rendement du ministère de l'Éducation de l'Alberta pour le curriculum d'études sociales de 3e année :

> Voici le texte de la question nº 7 d'une série de neuf questions modèles : « Quelle image illustre LE MIEUX le respect d'une coutume? ». Quatre images se trouvent sous la question — une camionnette, trois Autochtones vêtus de leur costume traditionnel, des enfants jouant au hockey en uniforme et un groupe d'adolescents qui dansent. Les trois dernières images représentent toutes le « respect d'une coutume ». Or,

> une seule image correspond à la « bonne réponse ». Par conséquent, un enfant de 3e année qui choisirait correctement l'une des deux autres images serait noté comme ayant donné une « mauvaise réponse »; tout raisonnement ou toute argumentation au sujet de ce qui s'avère exact est rejeté d'emblée même si cette argumentation ou ce raisonnement est entièrement justifié; et une intelligence véritablement critique sera désorientée par un test manifestement arbitraire dans son choix des bonnes et des mauvaises réponses★ (p. 95).

Dans les écoles étatsuniennes, l'emploi de plus en plus généralisé de « mesures de responsabilisation par les tests »★ s'accompagne d'un déclin de méthodes d'enseignement favorisant des aptitudes d'apprentissage d'ordre supérieur comme la rédaction de dissertations, les projets de recherche, les travaux de laboratoire et les discussions menées par les élèves (Darling-Hammond, 1991, p. 222).

Puisque les tests standardisés sont conçus pour classer les élèves en groupes, les questions sont choisies en fonction de leur tendance à élargir les résultats, et non en fonction de leur rapport avec le curriculum ou de leur validité prédictive.

Dans le cas des tests normatifs, la dispersion des résultats est nécessaire à l'établissement de comparaisons. On évite donc les questions qui ne contribuent pas à cette dispersion. Robertson (1998b) vulgarise ce concept plutôt technique :

> On exclut volontairement des tests standardisés les notions les plus couramment enseignées et apprises, non pas en raison d'une infâme conspiration mais à cause des besoins statistiques de l'analyse des données. Les questions qui assurent une meilleure distinction entre les élèves sont celles auxquelles 50 p. 100 des élèves seulement répondent correctement. Lorsque plus de 60 p. 100 des élèves répondent correctement à une

> question, celle-ci est systématiquement exclue. Or, même si ce genre de question est peut-être trop facile, elle porte aussi sans doute sur un domaine de contenu essentiel qui a été enseigné avec soin à presque chaque élève. Pour qu'ils donnent les résultats attendus, les tests standardisés doivent évaluer les notions périphériques et non les notions de base. Obligé de préparer les élèves à l'examen, le personnel enseignant doit se concentrer sur un contenu moins important — le genre de faits utiles qui conviennent aux examens à devinettes multiples — plutôt que sur des idées fondamentales ou des notions complexes★ (p. 71).

La réussite à un test est fonction de caractéristiques personnelles sans lien avec la connaissance de la matière.

Les pressions exercées pour obtenir de meilleurs résultats aux tests standardisés et « une mauvaise compréhension de la complexité et des causes de la réussite » se sont traduites par ce que l'on appelle la contamination des résultats des tests, c'est-à-dire les « facteurs influençant l'exactitude de l'interprétation des résultats »★ (Haladyna, Nolen et Haas, 1991, p. 4). La contamination des résultats se manifeste par diverses pratiques de préparation et d'administration des tests visant à augmenter les notes obtenues mais sans les relier à la matière apprise. Haladyna, Nolen et Haas (1991) en décrivent certaines et s'interrogent sur leur valeur sur le plan de l'éthique :

√ enseigner des techniques pour passer des tests;

√ élaborer un curriculum et des objectifs d'enseignement en fonction des tests;

√ présenter aux élèves des questions semblables à celles qui figureront dans les tests;

√ présenter à l'avance aux élèves les questions mêmes qui figureront dans les tests;

- √ utiliser du matériel pédagogique commercial conçu spécifiquement pour améliorer les résultats aux tests;

- √ le jour des tests, donner congé aux élèves dont le rendement est faible;

- √ influencer les réponses en donnant, par exemple, des indices ou des réponses aux élèves, ou encore en modifiant leur feuille de réponse.

Parmi les autres facteurs pouvant influencer les résultats, mentionnons le niveau d'anxiété et de motivation des élèves (Haladyna, Nolen et Haas, 1991). Il a été prouvé que les élèves anxieux obtiennent de moins bons résultats que ceux qui sont détendus, particulièrement dans le cas des questions difficiles. D'après une méta-analyse de plus de 500 études effectuée par Hembree (cité dans FairTest, 1995), « les élèves les plus anxieux à l'idée de passer un test ont moins d'estime de soi que ceux qui sont moins anxieux [...] Le niveau d'anxiété est plus élevé chez les membres du sexe féminin que chez les membres du sexe masculin, chez les élèves noirs que chez les élèves blancs dans les écoles élémentaires, et chez les élèves hispanophones que chez les élèves blancs de tous les âges »* (p. 9).

Selon Meaghan et Casas (1995d), les résultats aux tests standardisés « étant très sensibles aux facteurs motivationnels [..., ils] ne mesurent pas très bien les connaissances »* (p. 47). Les facteurs motivationnels ont fait l'objet d'une étude sur les résultats obtenus par des élèves au volet mathématiques de l'Iowa Test of Basic Skills. Plus de 400 élèves de l'élémentaire ont été divisés au hasard en deux groupes. Ceux qui faisaient partie du groupe expérimental ont reçu des directives spéciales de leurs enseignantes et de leurs enseignants sur l'importance de faire de leur mieux pour réussir le test et ont obtenu des résultats considérablement supérieurs à ceux du groupe de contrôle, qui n'avaient reçu aucune directive (Meaghan et Casas, 1995c, p. 42-43).

Le temps passé en classe à préparer les élèves et à administrer les tests est précieux et pourrait être employé à l'enseignement.

Dans certains cas, les programmes d'études sont délaissés au profit de la préparation des élèves pour les tests, préparation qui commence parfois des mois à l'avance (Medina et Neill, p. 26). Un certain nombre d'enseignantes et d'enseignants ontariens qui ont préparé leurs élèves au test ontarien de lecture et d'écriture de 9e année ont déclaré en privé avoir consacré une grande partie de la fin de l'année scolaire 1993-1994 à des activités préparatoires (Meaghan et Casas, 1995b, p. 83). En 1997, l'organisme ontarien chargé de l'administration des tests, l'Office de la qualité et de la responsabilité en matière d'éducation, a effectué une évaluation provinciale de tous les élèves de 3e année en compétences langagières et en mathématiques (ce test annuel administré à l'échelle de la province a depuis été étendu à tous les élèves de 6e année). Robertson (1998b) et Ireland (1997a) ont documenté certaines des limitations de cet exercice, dont l'administration a demandé dix journées complètes et a coûté plus de sept millions de dollars. Le temps requis pour l'administration du test aux élèves de 3e année a été réduit l'année suivante mais nécessitait tout de même trois heures par jour pendant cinq jours. Des questions éthiques se posent lorsque l'on consacre à des activités de testage, dont l'utilité est restreinte, un temps précieux qui devrait être réservé à l'enseignement et à l'apprentissage.

Le personnel enseignant est amené à enseigner en fonction des tests plutôt qu'en vue de l'apprentissage; en conséquence, les programmes d'études sont de plus en plus axés sur les tests (Meaghan et Casas, 1995d, p. 47).

Les comptes rendus de recherche décrivent très bien cette réalité. De nombreux enseignants et enseignantes ont dit qu'ils adaptaient leur programme d'études au contenu des tests, mais qu'ils s'inquiétaient du fait que les tests standardisés limitaient le programme d'études et qu'ils se sentaient obligés d'améliorer les notes obtenues (Meaghan et Casas, 1995c, p. 38-39). Par exemple, selon une enquête nationale menée auprès de la profession enseignante par le Boston College Center for the Study of Testing,

Evaluation, and Educational Policy, les enseignantes et les enseignants ayant dans leur classe un pourcentage important d'élèves appartenant à des minorités (plus de 60 p. 100) étaient davantage portés à enseigner en fonction des tests standardisés — en mettant l'accent sur les compétences de base et en consacrant plus de temps à des activités relatives aux tests — que leurs collègues qui enseignaient surtout à des élèves blancs (Madaus, 1994, p. 83-84). Wideen et autres (1997) ont étudié l'incidence des examens de fin d'études de 12e année sur l'enseignement des sciences dans plusieurs districts scolaires de Colombie-Britannique. Ils ont conclu que ces tests à grands enjeux nuisaient à la variété des approches pédagogiques utilisées, particulièrement en 12e année, parce qu'ils « dissuadaient les enseignantes et les enseignants d'utiliser des stratégies qui encourageaient l'enquête et l'apprentissage actif des élèves »; en outre, « cet appauvrissement avait une incidence sur la langue employée en classe »* (p. 428).

Autres problèmes

Le recours excessif aux tests standardisés cause d'autres problèmes :

√ on néglige les services éducatifs essentiels au profit de programmes de tests standardisés coûteux (et discutables), en allouant à ceux-ci des ressources financières déjà limitées, à une époque où les gouvernements imposent de fortes compressions au secteur de l'éducation (Meaghan et Casas, 1995b, p. 87); l'administration de tests à tous les élèves ontariens des 3e et 6e années devrait coûter environ 12 millions de dollars en 1999;

√ on ne tient pas compte des variations entre les curriculums et leur organisation pour les élèves d'une même année scolaire dans différents pays lorsque l'on compare les résultats des tests internationaux (Ireland, 1995);

- √ les tests standardisés ne fournissent aux éducatrices et éducateurs aucune information qui puisse les aider à reconnaître et à améliorer les programmes scolaires inefficaces (Casas et Meaghan, 1995, p. 19);

- √ la responsabilité des curriculums n'est plus confiée à la profession enseignante ni aux autorités scolaires, mais à l'industrie du testage et aux autorités gouvernementales, ce qui contribue à réduire la responsabilisation plutôt qu'à la renforcer (Meaghan et Casas, 1995d, p. 47).

Aux États-Unis, la production de tests standardisés relève principalement de maisons d'édition commerciales et d'organismes autres que les écoles (Darling-Hammond, 1991, p. 220). La situation est la même au Canada — les tests standardisés sont conçus et commercialisés par les plus importantes maisons d'édition de manuels scolaires et les profits sont versés à des sociétés mères étatsuniennes (Barlow et Robertson, 1994, p. 120). L'industrie privée du testage génère chaque année aux États-Unis des milliards de dollars et, en dépit de son influence croissante sur le système scolaire, elle n'est pratiquement ni réglementée ni responsable devant qui que ce soit. Dans de telles conditions, toute participation accrue du personnel enseignant et des parents au processus décisionnel en matière d'éducation est exclue (Medina et Neill, 1990).

Les tests et l'iniquité en éducation

Non seulement les tests standardisés sont difficiles à justifier d'un point de vue pédagogique, mais ils peuvent franchement être nuisibles aux élèves les plus vulnérables. L'emploi de ces tests est loin d'être neutre; il perpétue et intensifie les iniquités scolaires de deux façons : par l'utilisation abusive des résultats et par la partialité des tests qui défavorise les élèves provenant de groupes à faible revenu, les minorités raciales et ethniques, les filles et les jeunes femmes ainsi que les élèves handicapés. Darling-Hammond (1994) nous rappelle que des « recherches approfondies ont été faites sur le rôle du testage dans le renforcement et l'élargissement des inégalités sociales dans le domaine de l'éducation, [...] rôle qui a été largement reconnu »* (p. 10).

Mauvaise utilisation des tests

Les tests sont souvent utilisés de façon peu judicieuse. Au cours des ans, de nombreux tests mis au point dans un but précis ont été utilisés à d'autres fins. Le Scholastic Aptitude Test (SAT) (test d'aptitude scolaire) en est un parfait exemple. Conçu à l'origine pour évaluer les possibilités de réussite des élèves au cours de leur première année d'université aux États-Unis, le SAT a servi de critère pour l'admissibilité aux équipes sportives et aux prêts étudiants, ainsi que pour l'octroi de bourses d'études (Madaus, 1991, p. 227). Selon Schwartz (1999), en dépit de ses limitations (y compris sa « valeur prédictive limitée »), le SAT « est incontestablement devenu le test le plus important pour les élèves du secondaire aux États-Unis — un rite de passage universitaire et psychique qui influe profondément

sur leur choix d'orientations éducationnelles, provoque une vive anxiété et constitue pratiquement un baromètre mystique de l'estime de soi »* (p. 30).

Les tests administrés à des élèves individuels servent à analyser la réussite des politiques, des programmes, des écoles ainsi que du personnel enseignant et, de plus, ils sont abusivement utilisés comme « gardiens de l'éducation » et comme fondement d'importantes décisions concernant les élèves, les enseignantes et les enseignants, les écoles et le système scolaire dans son ensemble (Medina et Neill, p. 24). Les résultats des tests standardisés deviennent le mécanisme qui permet d'adopter un certain nombre de pratiques éducatives douteuses contribuant à l'iniquité en éducation.

Répartition en classes homogènes

Aux États-Unis, on se sert systématiquement des résultats de tests pour prendre d'importantes décisions concernant le placement des élèves. Les enfants provenant de milieux socioéconomiques désavantagés se trouvent ainsi regroupés dans des classes et programmes de niveau moyen et faible. Pourtant, la plupart des éducatrices et éducateurs progressistes croient que la répartition en classes homogènes contribue à créer une image de soi négative et à réduire les attentes à l'égard de l'avenir des élèves. Cette façon de procéder prive aussi les élèves placés dans les classes de faible niveau des avantages de travailler aux côtés d'élèves doués. Au Canada, les tests standardisés risquent d'être utilisés pour susciter un regain d'intérêt envers les classes homogènes, ce qui nous éloigne d'un modèle d'éducation plus inclusif (Meaghan et Casas, 1995a). Les écarts que l'on observe entre le rendement des élèves de race blanche aux États-Unis et celui des élèves issus de groupes minoritaires, ainsi qu'entre le rendement des élèves provenant de groupes à revenu élevé et celui des élèves de groupes à faible revenu est en grande partie attribuable à la répartition en classes homogènes (Darling-Hammond, 1991, p. 222). Peut-être parce que le programme d'études des classes de faible niveau tend à être limité et à mettre l'accent sur la mémorisation, les élèves de ces classes « faibles » n'obtiennent pas d'aussi bons résultats que leurs pairs d'aptitudes

comparables placés dans des programmes de formation générale ou non homogènes. Selon Oakes (1985), la répartition en classes homogènes nuit aux enfants issus de familles à faible revenu et de groupes minoritaires; non seulement elle reflète les inégalités de classe et de race présentes dans la société, mais elle les perpétue. Employer des tests biaisés envers certains groupes d'élèves pour ensuite placer ceux-ci dans des programmes de niveau inférieur est le comble de l'iniquité.

Promotion des élèves et redoublement

Les résultats des tests standardisés servent également à établir si les élèves passeront à l'année supérieure ou redoubleront leur année. La pratique du redoublement peut contribuer à la hausse du taux de décrochage (Melina et Neill, p. 29) et, selon Shepard et Smith, « contrairement à ce que l'on croit généralement, redoubler une année n'aide *pas* les élèves à progresser sur le plan scolaire et a des répercussions négatives sur leur adaptation sociale et leur estime de soi. »* (cités dans Darling-Hammond, 1991, p. 222). La répétition d'une année aboutit également à un rendement inférieur. Comme les filles ont généralement une moins bonne estime d'elles-mêmes que les garçons, les incidences de cette politique méritent d'être sérieusement pesées car elles risquent de nuire davantage encore à l'estime de soi des filles et des jeunes femmes.

Diplomation

Les résultats des tests servent également à déterminer l'admissibilité des élèves au diplôme d'études secondaires. Selon Darling-Hammond (1994) :

> Cette pratique repose sur le principe voulant que pour obtenir leur diplôme, les élèves doivent montrer qu'ils maîtrisent les « compétences minimales » requises pour travailler ou pour poursuivre leurs études. On présume que les tests évaluent adéquatement les compétences en question. Même si cette hypothèse est plausible en théorie, il est fort peu probable qu'elle le

> soit en réalité étant donné l'écart qui existe entre les tests à choix multiples portant sur des fragments d'information hors contexte et les exigences d'emplois et de tâches véritables d'adultes [...] *En fait, les recherches indiquent que ni l'employabilité ni les gains ne sont influencés de façon appréciable par les résultats des élèves aux tests de compétences fondamentales, alors que la possibilité d'obtenir un emploi et les risques de dépendance à l'égard de l'aide sociale sont étroitement liés à l'obtention du diplôme d'études secondaires* [les italiques ajoutés] [...] Par conséquent, l'emploi de ces tests comme seul facteur déterminant de l'obtention du diplôme est lourd de conséquences pour la personne et pour la société et ne comporte aucun avantage social évident* (p. 14).

À l'heure actuelle, 19 états américains obligent les élèves à réussir des tests d'état pour obtenir leur diplôme d'études secondaires (Education Week on the Web, 1999). La majorité des provinces canadiennes exigent actuellement que les élèves passent des examens de fin d'études secondaires comptant pour 30 à 50 p. 100 de leur note finale (Dunning, 1997). Alors que le diplôme d'études secondaires était autrefois la preuve d'un développement cumulatif sur un certain nombre d'années, il pourrait éventuellement être remplacé par un seul test à grands enjeux.

Autres exemples d'emploi abusif des tests

Au Canada, l'utilisation abusive des résultats des programmes d'évaluation mis en œuvre à l'échelle des provinces prend diverses formes, notamment « leur interprétation inadéquate par les politiciens et politiciennes et par les médias, et leur explication insuffisante au public (par exemple, on ne fournit aucune donnée générale et on ne compare pas d'échantillons présentant des caractéristiques différentes) »* (Gilliss, 1994, p. 6). D'après une enquête effectuée auprès des ministères de l'Éducation et d'un échantillon de conseils scolaires locaux dans l'ensemble du pays, la

mauvaise interprétation des résultats par le public et les médias constitue l'une des principales faiblesses des programmes provinciaux de testage (Traub, 1994, p. 1 et 3).

Lorsqu'il s'est avéré que 90 p. 100 des élèves de 9 e année ayant passé un test ontarien de lecture et d'écriture avaient des aptitudes à la lecture « satisfaisantes », résultats abondamment diffusés par les médias, le vice-président du conseil scolaire de North York « a annoncé que ces résultats n'étaient pas satisfaisants pour son conseil », malgré le fait

> qu'il n'avait aucunement participé à la préparation du test, qu'il ne disposait d'aucun renseignement sur la fiabilité des résultats ni sur les caractéristiques particulières des élèves de North York (dans la communauté urbaine de Toronto), notamment sur la proportion supérieure à la moyenne d'élèves d'ALS (anglais langue seconde)* (Meaghan et Casas, 1995b, p. 83).

Les résultats aux tests standardisés servent dans certaines administrations étatsuniennes de fondement aux décisions concernant les augmentations de salaire au mérite ainsi que l'obtention et le renouvellement du certificat d'aptitude pédagogique. De même, la répartition des fonds et l'accréditation et la désaccréditation des systèmes scolaires font partie des décisions touchant les écoles et les systèmes scolaires prises en fonction des résultats des tests (Medina et Neill, p. 24). Se fonder sur les résultats moyens obtenus par les écoles pour décider des récompenses et des sanctions à allouer aux enseignantes et enseignants et aux écoles, comme le font déjà certains états et districts scolaires aux États-Unis, est une façon d'inciter les systèmes scolaires à l'excellence en tenant compte des besoins du marché. George Bush, ancien président des États-Unis, proposait dans sa politique « America 2000 » que l'attribution des fonds gouvernementaux aux écoles soit fonction des résultats d'un test national. Cette politique encouragerait les écoles à fermer leurs portes aux élèves qui risqueraient de faire baisser

la moyenne des résultats, comme les enfants ayant des troubles d'apprentissage ou d'autres déficiences, ceux dont l'anglais n'est pas la langue première « ou ceux qui proviennent de milieux défavorisés sur le plan scolaire »* (Darling-Hammond, 1991, p. 223). De même, le personnel enseignant chevronné serait porté à éviter les écoles qui accueillent les enfants défavorisés, ce qui réduirait davantage encore les possibilités offertes à ces enfants.

Darling-Hammond (1991) cite une étude portant sur un important district scolaire urbain aux États-Unis, qui décidait des récompenses et des sanctions à attribuer aux écoles selon la moyenne des résultats qu'elles obtenaient aux tests : Darling-Hammond qualifie cruellement l'impact de cette façon de procéder d'« ingénierie généralisée des populations étudiantes ». Les auteurs de l'étude déclarent que :

> « la sélection des élèves est la meilleure façon de miser à court terme au jeu de la responsabilisation [...] Pour mettre toutes les chances de son côté, il suffit 1) d'ajouter quelques élèves de haut calibre, 2) d'éliminer certains élèves au rendement douteux et de garder ceux du milieu. L'admission des élèves est au cœur même de la responsabilisation »* (tiré de Darling-Hammond, 1991, p. 223).

Comme il est indiqué ci-dessus, la pression exercée pour hausser la moyenne des résultats aux tests peut se traduire par le rejet du système scolaire des élèves les plus à risque. Par exemple, les élèves obtenant de faibles résultats seraient placés dans des programmes d'enfance en difficulté afin d'éviter que leurs résultats ne paraissent dans les rapports de l'école; on refuserait d'accepter les élèves ayant de faibles notes dans les « écoles à admission ouverte »; et même, on encouragerait ces élèves à abandonner leurs études (Darling-Hammond, 1991, p. 223). De plus, on insiste tellement sur l'augmentation et le maintien de la moyenne des résultats aux tests

> qu'au fur et à mesure que le personnel enseignant, les directrices et les directeurs d'école et les surintendantes et surintendants se rendront compte que les résultats des tests ont une incidence cruciale sur leur carrière et sur la vie de leurs élèves, l'administration des tests et les curriculums se confondent. La diffusion des résultats des tests exerce une pression additionnelle sur le personnel enseignant et l'administration, qui réagissent aux nouvelles exigences en concevant des moyens ingénieux de hausser les résultats sans nécessairement améliorer les niveaux de compétence★ (Meaghan et Casas, 1995c, p. 49).

Comme le faisait remarquer Elliot Eisner au cours d'un exposé présenté à l'assemblée annuelle de la Fédération des enseignantes et enseignants de la Colombie-Britannique en mars 1998, « il est tout à fait possible pour le personnel enseignant d'une école d'améliorer les résultats des tests tout en diminuant la qualité de l'enseignement »★.

Responsabilité à grands enjeux axée sur la mesure

Selon une récente enquête réalisée par *Education Week* sur les politiques de responsabilisation dans 50 états américains (publiée sous le titre fort éloquent « Quality Counts '99: Rewarding Results, Punishing Failure » (la qualité par-dessus tout en 1999 : récompenser les réussites, punir les échecs »★), la responsabilisation en fonction des résultats est fortement encouragée pour promouvoir la réforme de l'éducation aux États-Unis. Le rapport révèle que les éléments (désignés « étapes essentielles ») nécessaires à la mise en œuvre d'un système complet de responsabilisation comprennent entre autres les tests, les normes, les bulletins scolaires, les systèmes d'évaluation, et les récompenses et sanctions. En fondant son analyse sur ces événements, l'enquête révèle que :

√ Quarante-huit états ont en place des programmes de textes dans l'ensemble de l'état (13 d'entre eux n'utilisent que des tests à choix multiples).

√ Trente-six états publient une fois par an des « bulletins de notes » sur le rendement de chaque école.

√ Dix-neuf états évaluent le rendement des écoles, y compris les écoles à « faible rendement », et publient les cotes d'évaluation.

√ Seize états ont le pouvoir de fermer, de prendre en main ou de « reconstituer » les écoles à rendement insatisfaisant (dans le rapport, « reconstitution » désigne le « remplacement total ou quasi total du personnel de l'école et sa réouverture avec une nouvelle direction »★). Il n'est pas étonnant que les écoles fréquentées par les élèves pauvres ou appartenant à des groupes minoritaires soient les plus susceptibles de subir des sanctions.

√ Quatorze états offrent des récompenses pécuniaires aux écoles en fonction de leur rendement.

√ Deux états ont tenté de lier l'évaluation du personnel enseignant au rendement des élèves.

Le rapport met en opposition cette « méthode dure de responsabilisation » et une « méthode humanisée » — qui consiste à fournir aux écoles et aux élèves des ressources, un appui et un encouragement suffisants, ainsi qu'à encourager la responsabilité professionnelle. Le concept de la responsabilisation professionnelle tient compte du lien critique entre la qualité du personnel enseignant et le rendement des élèves et, partout, de l'importance d'investir dans l'épanouissement et le perfectionnement professionnel des enseignantes et des enseignants.

En mars, le gouvernement de l'Alberta annonçait l'adoption de sa propre version de la responsabilité axée sur les résultats pour les classes de la maternelle à la 12e année. Selon un communiqué de

100

presse du ministère de l'Éducation de cette province (11 mars 1999), le ministère entend consacrer 66 millions de dollars à un programme d'incitation à l'amélioration du rendement scolaire, qui se veut « un projet innovateur conçu pour améliorer l'apprentissage des élèves », en accordant des fonds supplémentaires aux conseils scolaires « selon la mesure dans laquelle ils ont atteint des objectifs d'amélioration ». Des objectifs seront établis par la province de concert avec les conseils scolaires locaux. La répartition des fonds supplémentaires (se situant entre 2 et 4 p. 100 du budget des salaires d'un conseil) dépendra de l'amélioration de l'apprentissage évaluée à partir des résultats obtenus au test de rendement provincial et à l'examen de fin d'études secondaires, du taux de diplomation des écoles secondaires et « d'indicateurs de rendement établis localement »★, comme l'assiduité des élèves et les résultats des tests de rendement locaux. L'adhésion au programme est volontaire. Celui-ci (apparemment sans précédent au Canada) sera mis en œuvre au cours de l'année scolaire 2000-2001. Comme il a été mentionné plus haut, les États-Unis s'y connaissent davantage en ce domaine — en effet, 14 états, dont la Caroline du Nord, le Maryland et le Texas, ont institué une forme quelconque de programme d'encouragements financiers (« Alberta Proposes Bonuses for Student Achievement Gains », 1999).

L'Alberta Teachers'Association (ATA) craint qu'un programme qui offre essentiellement des primes en comptant en échange de meilleurs résultats dans les tests, « ne mette les écoles et les conseils en concurrence et *ne favorise* injustement les districts aisés, surtout en cette époque de compressions budgétaires. »★ (McMahen, 1999). Le programme proposé d'incitation à l'amélioration du rendement scolaire soulève d'autres questions : Quel effet aura-t-il sur la collégialité au sein du personnel enseignant? Le programme poussera-t-il un nombre accru d'enseignantes et d'enseignants à demander des mutations à des écoles où le rendement des élèves est plus élevé? Quelles sont les retombées éventuelles de l'identification d'écoles et de salles de classe n'atteignant pas les objectifs cibles? (« Alberta Proposes Bonuses for Student Achievement Gains », 1999) Affirmant que de tels programmes d'incitation sont incompatibles avec ce que l'on connaît des pratiques efficaces

d'évaluation de l'enseignement et des élèves, de la motivation du personnel enseignant et de l'amélioration des écoles en général, l'ATA et ses partenaires éducationnels de la province ont proposé une solution de rechange destinée à améliorer l'enseignement et l'apprentissage à la lumière de recherches et de pratiques éprouvées. Le ministre de l'Éducation a convenu de mettre le programme en veilleuse dans l'attente de l'étude du programme de rechange (« Improvement, Not Incentives », 1999).

La méthode dite dure telle qu'elle est décrite dans le rapport d'enquête d'*Education Week* n'est manifestement pas la voie à suivre pour favoriser la responsabilité véritable en éducation, quoi qu'en disent ses défenseurs et défenseures. Darling-Hammond (1994) explique les effets nuisibles d'une telle approche sur l'équité :

> Il va sans dire que ce genre de politique, qui récompense ou punit les écoles selon les points totaux obtenus aux tests, crée une vision déformée de la responsabilisation, vision qui incite à manipuler les chiffres en jouant avec le placement des élèves, au détriment des efforts déployés pour répondre efficacement aux besoins éducationnels de ces derniers [...] L'application de sanctions aux écoles obtenant de faibles résultats aux tests pénalise doublement les élèves défavorisés : en plus de les cantonner dans des écoles médiocres, la société les punit une nouvelle fois parce qu'ils ne réussissent pas aussi bien que les élèves qui fréquentent des écoles dotées de meilleures ressources et d'un personnel enseignant plus compétent. Ce genre de régime de récompenses confond la qualité de l'éducation qu'offrent les écoles et les besoins des élèves qui les fréquentent; il nuit à l'équité et à l'intégration, et écarte toute possibilité d'offrir de façon juste et ouverte un choix d'écoles en dissuadant les bonnes écoles d'ouvrir leurs portes aux élèves défavorisés sur le plan scolaire. Un tel régime de récompenses accorde plus d'importance à la manipulation des résultats et à

> l'affectation ou à l'exclusion des élèves qu'à l'amélioration des écoles et à l'élaboration de méthodes d'enseignement plus efficaces★ (p. 15 et 16).

Dans le contexte de l'adoption de tests nationaux envisagée aux États-Unis, Madaus et Kellaghan (1993) analysent l'emploi des tests comme moyen de déterminer les politiques éducationnelles et rapportent que « le lien entre les résultats obtenus aux tests et les importantes récompenses ou sanctions qui en découlent, est l'élément clé qui fait que les tests ont une influence puissante sur les politiques » — d'où l'expression « tests à grands enjeux ». Dans un article intitulé « Testing as a Mechanism of Public Policy : A Brief History and Description », les auteurs se servent de six principes de base pour décrire les processus « selon lesquels les tests utilisés pour déterminer les politiques touchent la vie des personnes et des institutions ».★ Leurs observations sont perspicaces. (tiré de Madaus et Kellaghan, 1993, p. 8 et 9) :

1. Les tests et les examens influencent les personnes, les institutions, les curriculums et l'enseignement selon la façon dont ils sont perçus — c'est-à-dire, si les élèves, le personnel enseignant ou l'administration croient que les résultats d'un examen sont importants, il importe très peu qu'ils aient raison ou non — l'effet produit dépend de la perception de chacun.

2. Si l'on croit que des décisions importantes sont liées aux résultats des tests, les enseignantes et enseignants enseigneront en fonction des tests.

3. Partout où des tests à grands enjeux sont en place, une tradition de questions antérieures se forme, et cette tradition en vient à définir en réalité le curriculum.

4. Lorsque les résultats des tests font partie des critères ou sont le seul critère déterminant les choix d'études et de vie des élèves, la société tend à faire de ces résultats le but premier de la scolarité plutôt qu'un indicateur utile, mais imparfait, du rendement.

5. Les tests à grands enjeux donnent aux organismes qui les préparent ou les contrôlent toute autorité sur les curriculums.

6. Un principe très général s'applique à tous les indicateurs sociaux importants, y compris aux tests, et met en lumière l'orientation non intentionnelle et très souvent imprévue donnée aux enjeux. Ce principe repose sur l'idée que plus un indicateur social quantitatif est utilisé pour prendre des décisions d'ordre social, plus cet indicateur risque de dénaturer et de corrompre le processus social qu'il vise à contrôler.*

Même les méthodes d'évaluation authentique des élèves, utilisées en remplacement des tests standardisés, ne sont pas à l'abri de ces conséquences. Il semblerait que lorsque l'évaluation authentique (étudiée de façon plus détaillée dans une autre section de ce rapport) est liée à une responsabilité dont les enjeux sont importants, c'est au détriment de ses avantages pédagogiques (Jones et Whitford, 1997; Torrance, 1993). Par exemple, selon le système de responsabilisation par des tests à grands enjeux du Kentucky, les résultats des évaluations fondées sur le rendement sont convertis en une « note d'école » et l'on a largement recours aux récompenses et aux sanctions externes. D'après Jones et Whitford (1997),

> si la note d'une école dépasse les attentes de l'état pour cette école, le personnel enseignant et l'administration peuvent recevoir des primes substantielles. Si la note n'est pas assez élevée ou cesse d'augmenter avec les années, le personnel enseignant et l'administration peuvent être soumis à une période de probation, et l'école peut même passer sous le contrôle de l'état* (p. 276).

La décision de l'état d'élaborer, de façon arbitraire et sans données empiriques, la formule qui servirait à déterminer les notes que les écoles doivent obtenir (et pour mesurer l'amélioration de chaque école) a particulièrement irrité la profession enseignante et les administratrices et administrateurs.

La mise sur pied du système informatisé de résultats scolaires du Kentucky (KIRIS) a notamment eu pour effet d'accroître le contrôle de l'état sur le curriculum aux dépens des autorités locales et du jugement professionnel des enseignantes et des enseignants. Comme l'expliquent Jones et Whitford,

> la pression exercée par l'administration de tests de l'état comme moyen de responsabilisation a incité le personnel enseignant des écoles à exiger plus de détails sur le contenu à évaluer. Par le fait même, le pouvoir décisionnel des écoles locales en matière de curriculum en a été amoindri. Cette formule dialectique contribue à accroître le contrôle de l'état sur le curriculum local* (p. 278).

Au lieu d'obtenir le résultat peut-être souhaité, soit un enseignement constructiviste et axé sur le rendement, le KIRIS a mis l'accent sur un « apprentissage qui peut être mesuré de façon fiable et valide », autrement dit, l'évaluation de résultats observables et quantifiables à l'aide de tests définit effectivement les connaissances qu'il vaut la peine d'acquérir. L'évaluation du rendement peut certes améliorer l'apprentissage des élèves, mais ce potentiel est corrompu par l'emploi de tests à grands enjeux. Les auteurs concluent que les principes sous-jacents à la responsabilisation et à l'évaluation du rendement dans un tel contexte sont fondamentalement incompatibles : la responsabilisation « encourage la conformité à des normes imposées de l'extérieur, tandis que [l'évaluation du rendement] résulte d'interactions émergentes entre le personnel enseignant et les élèves dans des classes locales dynamiques »* (Jones et Whitford, p. 280).

La partialité des tests

Les tests standardisés montrent d'ordinaire un certain parti pris envers l'un ou l'autre sexe.[5] Rosser et autres (1992) remarquent que les jeunes femmes obtiennent immanquablement des notes plus élevées au secondaire *et* à l'université aux États-Unis, mais qu'elles ont les plus faibles notes au Scholastic Aptitude Test (SAT) et aux autres tests semblables censés prédire le rendement scolaire des élèves.[6] Au secondaire, les filles qui souhaitent poursuivre des études universitaires représentent plus de 52 p. 100 des élèves qui passent ces tests. Toutefois, ces examens sous-estiment systématiquement leurs aptitudes, et leurs chances de fréquenter des établissements

[5] Selon Robertson (1992), « les différences de résultats entre garçons et filles ne prouvent pas forcément une partialité des tests. N'oublions pas qu'il est possible qu'un groupe en sache réellement davantage sur un sujet qu'un autre groupe. Tout gauchissement est donc le résultat d'un type de connaissances que l'on valorise plutôt que d'une partialité en faveur ou au désavantage d'un groupe donné. Cependant, si un test ne répond pas au critère de validité, c'est-à-dire si les groupes en savent autant sur le même sujet mais que l'un d'eux obtient constamment de meilleures notes à une épreuve visant à évaluer ces connaissances, ou si un test ne révèle pas le niveau de connaissances supérieur d'un groupe par rapport à l'autre, on peut dire que le test est entaché de partialité » (p. 26).

[6] Même le plus important fournisseur privé de tests éducatifs dans le monde, l'Educational Testing Service (ETS), (qui conçoit et administre le SAT), avoue que le SAT est partial en ce sens qu'il surestime le rendement prévu des garçons à l'université et sous-estime celui des filles. L'ETS reconnaît également que les notes du secondaire sont plus utiles que leurs tests pour prévoir la réussite à l'université (Thurmond, 1994). Dans un jugement historique rendu par la cour fédérale des États-Unis au sujet de la nature discriminatoire du SAT, le juge a déclaré que « [...]a probabilité que les femmes obtiennent au SAT une note immanquablement inférieure de 60 points à celle des hommes est pratiquement nulle, mis à part les facteurs discriminatoires ». Même lorsqu'on tient compte de facteurs comme l'origine ethnique et la scolarité des parents, il subsiste un écart considérable entre les notes des filles et celles des garçons (Rosser et autres, 1992, p. i et ii).

postsecondaires, d'obtenir des bourses d'études et d'être admises à des programmes spéciaux pour élèves doués s'en trouvent réduites. Les élèves qui obtiennent de faibles notes au SAT ont des attentes réduites par rapport à leurs capacités, ce qui influence leur choix des universités auxquelles ils présenteront une demande d'admission (Weiss, 1991, p. 108). À plus long terme, sur les plans économique et personnel, les femmes auront moins de chances de se trouver un bon emploi, toucheront des revenus moins élevés et auront moins confiance en elles-mêmes. Les filles provenant de familles à faible revenu font face à une double discrimination dans les tests, soit en raison de leur sexe et en raison de leur classe sociale, et sont effectivement pénalisées deux fois (Rosser et autres, 1992). En somme, les résultats du SAT sont davantage révélateurs de facteurs comme le revenu familial, la race et le sexe que de la capacité des élèves de réussir à l'université (Robertson, 1998b, p. 80).

Robertson (1992) étudie l'impact du testage sur les filles dans le contexte de l'enseignement des sciences, des mathématiques et de la technologie (SMT). Elle note que « les indicateurs complexes de la "réussite" des élèves et de l'enseignement sont souvent et à tort réduits à des chiffres tirés d'un seul test » (p. 25) et que cette orientation donnée à la réforme scolaire est fortement politisée. Chez les élèves du secondaire, des différences minimes dans les résultats obtenus aux tests suffisent pour déterminer non seulement qui réussira à un cours, mais aussi qui sera admis à des universités renommées (p. 25). Comme le financement insuffisant du système postsecondaire réduit les possibilités d'accès aux études postsecondaires, les universités canadiennes ont dû hausser leurs droits de scolarité et diminuer le nombre de places disponibles. Ces facteurs contribuent tous à réduire l'accès des femmes aux études supérieures, particulièrement dans les domaines (comme les SMT) où elles sont déjà sous-représentées.

Les conclusions ci-dessous (tirées de Robertson, 1992, p. 26 et 27) sont extraites du rapport *How Schools Shortchange Girls* (1992) publié par l'American Association of University Women (l'AAUW). Cette étude importante fait la synthèse des recherches effectuées sur

l'expérience scolaire des filles et conclut qu'il y a inégalité entre les filles et les garçons en ce qui concerne la quantité et la qualité de leur éducation :

√ Bien que les filles obtiennent de meilleures notes dans le secondaire et au collège, on a constaté que les garçons ont deux fois plus de chances qu'elles de bénéficier d'une bourse d'études dont l'attribution repose sur les résultats obtenus aux tests standardisés des écoles secondaires.

√ Les filles réussissent constamment mieux que les garçons aux tests d'aptitude oraux lorsqu'il s'agit de concepts et d'idées plus que de choses, ou d'esthétique, de philosophie, de relations et de tâches stéréotypées comme étant féminines. Les garçons obtiennent de meilleures notes aux tests oraux dans les domaines scientifique et sportif. Depuis les révisions apportées en 1978 pour augmenter le nombre des sujets à teneur scientifique dans les tests oraux, les notes intermédiaires des garçons en lecture sont supérieures de douze points à celles des filles, alors qu'elles ne l'étaient que de trois points auparavant.

√ Les résultats des tests d'aptitude scolaire sous-évaluent la réussite future des filles et surévaluent celle des garçons au collège et à l'université, notamment en mathématiques.

√ Les tests peuvent se révéler partiaux de différentes manières (contenu, présentation et matière). Ainsi, en mathématiques, les filles s'en sortiront plutôt mieux en calcul, en algèbre et en logique, alors que pour les garçons, ce sera en problèmes à résoudre et en géométrie. Les épreuves où l'accent est mis davantage sur une catégorie d'aptitudes que sur d'autres créeront des différences entre garçons et filles.

√ Les filles obtiennent de meilleurs résultats en dissertation et sur des questions à réponse libre; les garçons, aux questions à choix multiples. Les filles répondent à moins de questions que

> les garçons, sont moins nombreuses à terminer les tests et choisissent plus souvent « Je ne sais pas » quand elles en ont la possibilité.

FairTest signale que

> Le fait que le [SAT] soit chronométré et que l'on doive répondre à ses questions rapidement défavorise les jeunes femmes et les membres de groupes minoritaires. Deviner est également difficile pour les femmes, et peut-être pour les personnes de couleur. Le format du choix obligatoire ne permettant pas de nuances de sens, il désavantage les filles, dont le processus de raisonnement est plus complexe★ (Rosser et autres, 1992, Annexe I).

Dans un rapport intitulé *Gender Gaps : Where Schools Still Fail Our Children*, donnant suite au rapport *How Schools Shortchange Girls* et portant sur les progrès réalisés au chapitre de l'équité depuis le début des années 1990, l'Educational Foundation de l'AAUW (1998) conclut que les « tests à grands enjeux qui ont une influence disproportionnée sur la vie des élèves sont ceux qui font ressortir le plus radicalement les différences de rendement entre les sexes »★ (p. 35). Le rapport recommande entre autres que les établissements postsecondaires prennent des décisions sur l'admission des élèves et leur admissibilité aux bourses d'études à partir de « renseignements très variés », et que la recherche se concentre sur le rapport entre les notes des élèves et les résultats qu'ils obtiennent aux tests et leur sexe, leur race et leur origine ethnique, afin de fournir des renseignements « qui permettront de favoriser l'équité en tenant compte des différences individuelles » (p. 41).

Selon Robertson, même si les rapports de l'AAUW et les autres données disponibles ne peuvent prouver que le parti pris à l'égard d'un sexe désavantage systématiquement les filles dans les évaluations canadiennes en SMT, « il est curieux de voir combien on prête peu attention au risque apparent que représente pour les filles un passage

massif à des évaluations standardisées à réponses uniques et corrigées par ordinateur » (p. 27).

Aux États-Unis, la plupart des tests standardisés sont conçus pour et par une population blanche de classe moyenne à supérieure, et reflètent donc une optique particulière. Des études indiquent qu'un certain nombre de caractéristiques des tests standardisés peuvent fausser les résultats au détriment des élèves appartenant à des groupes minoritaires et de ceux issus de familles à faible revenu. Parmi ces caractéristiques, mentionnons (selon Medina et Neill, p. 8-10) :

√ Une partialité linguistique dans les tests de langue anglaise que passent les élèves dont la langue première n'est pas l'anglais; cette partialité se manifeste plus subtilement à l'égard des élèves parlant des dialectes anglais non standard lorsque les tests sont rédigés dans un style anglais plus formel et relevé, plutôt que dans la langue vernaculaire.

√ Les questions des tests ne tiennent pas compte des expériences, des points de vue et des connaissances culturelles des enfants provenant de groupes raciaux et ethniques minoritaires, de familles à faible revenu ou de quartiers urbains pauvres et de milieux ruraux.

√ Le style d'apprentissage et la résolution de problèmes diffèrent d'une personne à l'autre. Les différents styles d'apprentissage (et non les différentes aptitudes), souvent liés à des facteurs comme l'origine raciale et ethnique, le revenu et le sexe, ne sont pas pris en considération dans la conception des tests, parce qu'on suppose « que la façon de percevoir l'information et de résoudre des problèmes est la même pour tous et toutes ».★

Dans le cadre de leur recherche sur les tests d'aptitudes linguistiques administrés à des élèves noirs en Afrique du Sud, Peirce et Stein (1995) ont découvert que les interprétations multiples d'un même texte faisant partie d'un test de lecture étaient fonction de l'appartenance sociale des élèves. Les répercussions de leurs

conclusions débordent largement le contexte sud-africain en ce qui concerne l'iniquité en éducation. Selon les auteurs,

> essentiellement, les élèves qui proviennent de la même classe sociale, sont de la même race ou du même sexe que les personnes qui conçoivent les tests seront avantagés par rapport aux autres [...] Pour promouvoir l'équité dans l'évaluation scolaire, différents intervenants, par exemple des examinatrices et des examinateurs, des membres du personnel enseignant et de l'administration, des parents et des élèves, devraient pouvoir participer à l'élaboration des tests★ (p. 62).

Dans l'extrait ci-dessous, Weiss (1991) souligne à quel point il est injuste de se voir refuser l'accès aux études postsecondaires pour avoir obtenu de faibles résultats à un test entaché de partialité :

> Il est particulièrement frustrant pour les élèves appartenant à des groupes minoritaires, qui ont dû surmonter d'innombrables obstacles sociaux et économiques et qui ont obtenu de bons résultats scolaires au cours de leurs premières années d'études, de se voir refuser l'accès aux études supérieures parce qu'ils ont eu de faibles résultats aux examens d'admission à l'université ou à un établissement d'enseignement supérieur★ (p. 114).

Les cours de préparation à des tests comme le SAT sont devenus une industrie générant des millions de dollars. Les plus importantes entreprises de préparation au SAT sont les Princeton Review et Kaplan Educational Centers, ce dernier touchant des recettes annuelles de près de 300 millions de dollars (Schwartz, 1999, p. 32). Selon Weiss, ces cours peuvent tellement améliorer les résultats obtenus par des élèves au SAT qu'« au lieu de voir leurs demandes refusées, ils pourront être admis avec une bourse »★ (p. 118). Toutefois, les cours coûtent cher — un cours de six à huit semaines comprenant des exercices pratiques coûte environ 800 $, tandis que

les tutrices et tuteurs privés peuvent demander quelques centaines de dollars l'heure (Schwartz, p. 32) — et sont donc bien au-delà des moyens de la majorité des élèves. Ceux qui appartiennent à des groupes minoritaires et ceux qui proviennent de familles à faible revenu sont ainsi doublement désavantagés — selon Weiss, « [n]on seulement ils n'ont pas les moyens de bénéficier de cet encadrement, mais l'efficacité de la méthode accroît encore la disparité entre les groupes raciaux »★ (p. 118).

Le testage est devenu une pratique profondément ancrée dans notre société, et il ne se limite pas au domaine de l'éducation.[7] Certains disent que le testage est une forme de technologie sociale et soutiennent que, bien qu'il n'y ait rien de fondamentalement injuste à propos de la technologie en soi, « les technologies sont un produit de la structure actuelle des possibilités et des contraintes sociales », et qu'à ce titre, « elles sont susceptibles d'étendre, de façonner, de remanier ou de reproduire cette structure »★, y compris les iniquités qu'elle renferme (Madaus, 1994, p. 77 et 79). Autrement dit, les valeurs, les partis pris et les hypothèses de l'élite qui crée la technologie sont reproduits en elle. Parmi les valeurs sous-jacentes qui imprègnent les tests, mentionnons l'utilitarisme, la compétitivité économique, l'optimisme technologique, l'objectivité, le contrôle et la responsabilité bureaucratique, la précision numérique, l'efficacité, la normalisation et la conformité. De cette façon, précise Madaus (1994), « le testage influe non seulement sur des valeurs éducationnelles importantes comme la qualité de l'enseignement et

[7] Dans son livre, F. Allan Hanson étudie les hypothèses sociales et culturelles sur lesquelles repose l'administration des tests aux États-Unis. Il analyse le testage en tant qu'institution sociale si étroitement liée à la notion de méritocratie qu'on met rarement en question leur utilité et leurs autres présumées qualités. S'inspirant des travaux de Foucault, il suggère que les tests (qu'il s'agisse de tests de polygraphie et de consommation de drogues, ou de tests d'orientation professionnelle, d'intelligence et de rendement) nous définissent comme êtres sociaux et sont une forme de surveillance et de contrôle exercés sur les personnes par les organismes et les institutions qui utilisent et valorisent ces tests (Kleinsasser, 1994, p. 96 et 97).

les systèmes de récompense des élèves et du personnel enseignant, mais aussi sur le contenu et la méthode de l'enseignement et de l'apprentissage »* (p. 79).

Quelques mots sur l'évaluation authentique

L'évaluation authentique, ou évaluation du rendement, est souvent proposée comme solution de rechange aux tests standardisés. Selon FairTest, l'évaluation du rendement est « un terme général désignant une activité d'évaluation dans laquelle les élèves élaborent des réponses, créent des produits ou font des démonstrations pour témoigner de leurs connaissances et de leurs compétences »* (Neill et autres, 1997, Annexe B). Cette évaluation se fait au moyen d'initiatives très diverses, notamment des exposés oraux, des débats, des expositions, des recueils de rédactions d'élèves, des enregistrements vidéo de spectacles et d'autres expériences d'apprentissage, des constructions et des maquettes et leurs solutions à des problèmes, des expériences et des résultats de recherches scientifiques et autres, des observations faites par le personnel enseignant et des répertoires des travaux et des comportements des élèves, ainsi que des travaux coopératifs de groupe (Darling-Hammond, 1994, p. 5 et 6). [8] Cette forme d'évaluation relativement nouvelle pose quelques problèmes qui n'ont pas encore été réglés,

[8] Dans *Testing Testing: Social Consequences of the Examined Life*, de F. Allan Hanson, Kleinsasser (1994) fait remarquer que les résultats des tests standardisés, qui, vu leurs limites, ne témoignent qu'indirectement du rendement scolaire des élèves, sont néanmoins conservés par les autorités scolaires dans les dossiers scolaires officiels de ces derniers. Ces renseignements sont ensuite utilisés au moment où d'importantes décisions concernant les études et l'avenir des élèves sont prises; par contre, les travaux réalisés par les élèves (démonstration directe de leur rendement), qu'il s'agisse d'exemples de rédaction comme des dissertations ou des contes, de projets artistiques et d'expo-science, d'enregistrements audio et vidéo de musique et de spectacles de danse, etc., sont normalement conservés par les élèves (plutôt que par l'école) qui, parce qu'ils en sont les auteurs, leur attribuent une plus grande valeur. Toutefois, ces documents « perdent beaucoup de leur signification lorsqu'ils sont conservés par des autorités puissantes comme les écoles ou les employeurs [...] Lorsque les établissements valorisent les résultats de tests qui sont entre les mains d'autorités ayant le pouvoir de faire ou non des recommandations, ils en disent long aux élèves sur les valeurs et les critères d'évaluation de notre société »* (p. 98).

comme le temps considérable qu'exigent son élaboration et son administration (de par sa nature, l'évaluation authentique demande beaucoup de temps et de main-d'œuvre), les coûts qui s'y rattachent ainsi que les questions qui se posent sur sa validité et sa fiabilité.

Il ne suffit pas de modifier la forme et le contenu de l'évaluation (notamment de l'évaluation authentique) pour résoudre les problèmes associés à une évaluation plus traditionnelle. Darling-Hammond (1994) souligne que l'évaluation doit également être perçue en fonction de l'utilisation qui en sera faite, particulièrement pour appuyer et renforcer les projets favorisant l'équité dans le système scolaire. À son avis,

> *la modification de la forme de l'évaluation est peu susceptible d'améliorer l'équité si l'on ne modifie pas également les façons dont nous utilisons cette évaluation :* que ce soit comme mécanisme de classement ou comme support diagnostique; comme instrument externe de contrôle du rendement ou comme outil conçu à l'échelon local pour enquêter en profondeur sur l'enseignement et l'apprentissage; comme moyen d'appliquer des sanctions aux personnes déjà désavantagées ou comme levier permettant d'égaliser les ressources et d'accroître les occasions d'apprentissage★ (italiques dans l'original) (p. 7).

En même temps, il faut mettre en œuvre de plus vastes réformes qui prévoient le financement et le soutien adéquats des ressources sociales, familiales et scolaires et de santé, ce qui équivaut, selon Madaus (1994), à élaborer « des règles du jeu équitables pour les élèves et les écoles » en ce qui concerne les conditions sociales et scolaires. Il croit que

> les questions d'équité découlant de l'évaluation ne se limitent pas à l'instrumentation, à la technique ou à la psychométrie particulières au test. Plus généralement, ces questions se rattachent aux conditions sociales et scolaires du moment présent et du passé★ (p. 87).

Comme avec toute nouvelle technologie et l'aura qui l'entoure, l'évaluation authentique risque d'être perçue comme *la* solution aux problèmes de l'éducation, au point d'exclure complètement la considération d'autres facteurs comme la santé et les conditions de vie des élèves, la formation du personnel enseignant, etc. Madaus (1994) nous alerte au risque que les solutions technologiques « empêchent les décisionnaires et le public de se rendre compte que les États-Unis ne peuvent régler leurs problèmes d'éducation par des tests, des examens ou des évaluations »★ (p. 78). Sa mise en garde vaut certainement pour le contexte canadien, où l'on met de plus en plus l'accent sur la responsabilisation par les tests.

La politique des tests

Les tests standardisés doivent être étudiés dans une perspective politique autant que pédagogique. Préconisé comme une mesure du rendement scolaire et comme un moyen de régler nos problèmes d'éducation, l'emploi des tests standardisés est une méthode attrayante sur le plan politique, car tout en donnant l'illusion que l'on fait quelque chose pour améliorer la qualité de l'éducation (grâce à un mécanisme perçu comme étant objectif et neutre), il nous permet d'oublier la réalité complexe de l'éducation des enfants et des jeunes ainsi que les exigences, le temps et les coûts qui s'y rattachent. McMurtry (1992) constate que les tests standardisés « nous distraient parfois de façon néfaste de notre tâche la plus urgente en tant qu'éducatrices et éducateurs » et qu'« on peut en faire un usage politique abusif pour se soustraire aux exigences réelles de l'enseignement »★ (p. 95). Robertson (1998b) consacre un chapitre entier à la politique du testage dans son ouvrage *No More Teachers, No More Books: The Commercialization of Canada's Schools* et affirme « qu'il est beaucoup plus facile d'augmenter le nombre de tests que d'améliorer la qualité de l'enseignement, ou de remédier à l'inégalité des chances »★ (p. 72). Elle remarque également que les tests ont tendance à engendrer d'autres tests :

> Que les prétendues déficiences [dans les écoles] soient attribuables au curriculum, aux normes ou à la responsabilisation, la stratégie consiste à utiliser des données, quelque altérées qu'elles soient, pour faire avancer les réformes de l'éducation. On a donc recours à un mécanisme politique clé : l'emploi accru du testage — des tests plus nombreux, administrés plus souvent et à plus d'enfants, de façon à produire plus de données, qui sont diffusées à plus grande échelle et qui sont plus susceptibles d'être manipulées★ (p. 63 et 64).

Comparaisons internationales

En dépit du fait qu'il est difficile d'établir des comparaisons internationales du rendement scolaire, les données disponibles montrent que le Canada se classe près de la moyenne parmi les pays comparables et que dans certains cas, il obtient de meilleurs résultats (Rehnby, 1996, p. 15). Néanmoins, l'impression que nos écoles sont aux prises avec de sérieuses difficultés est attribuable en partie aux reportages sensationnalistes et trompeurs des médias sur les résultats canadiens aux tests standardisés internationaux. On peut effectivement induire les gens en erreur lorsqu'on prend comme indicateurs des problèmes de notre système scolaire les résultats aux tests internationaux. La méthodologie utilisée pour les tests internationaux de rendement en mathématiques et en sciences pose certains problèmes, notamment (d'après la Fédération canadienne des enseignantes et des enseignants, 1995a) :

- √ La participation aux études internationales est souvent volontaire et tous les systèmes scolaires n'y sont pas représentés. Les classements ne s'appliquent donc qu'aux pays qui ont bien voulu participer.

- √ Les pays participants procèdent différemment pour choisir les élèves qui passeront les tests (certains ne choisissent que les élèves les plus doués).

- √ Les pays participants ne sont pas évalués en fonction de leurs propres curriculums et on ne tente pas de rajuster les résultats des tests selon les conditions qui existent dans les divers pays.

Les études internationales antérieures ont clairement montré « que les données quantitatives peuvent être trompeuses à moins d'être accompagnées de renseignements qualitatifs détaillés pour les mettre en contexte » (Fédération canadienne des enseignantes et des enseignants, 1995a). Nagy (1995) fait remarquer que « les études internationales peuvent devenir des instruments utiles à la compréhension et à l'amélioration de l'éducation, mais qu'elles font

face à un dilemme. Au fur et à mesure qu'elles deviennent nécessairement plus raffinées, leur complexité même fait qu'elles se prêtent moins bien à une interprétation simpliste erronée. Or, elles risquent ainsi de perdre leur attrait politique et leur financement »★.

L'interprétation erronée d'études comparatives internationales complexes par le public et les milieux politiques peut déclencher un processus d'élaboration de politiques qui aboutira à des décisions mal fondées. Nagy (1996) établit une distinction importante entre les jugements de *principe* et les jugements *politiques* — les premiers sont des décisions éclairées prises à partir de rapports originaux, accompagnées de mises en garde au sujet de l'interprétation des données et d'analyses pertinentes d'universitaires et de spécialistes; les seconds concernent des déclarations sans fondement s'appuyant sur « l'interprétation de ces rapports par les médias et le milieu politique » (p. 1). Malheureusement, le second type de jugement est devenu trop courant.

Une présentation déformée des résultats des tests (que ce soit au niveau international, national ou provincial) contribue à augmenter les inquiétudes des parents, qui ont ainsi l'impression que l'éducation de leurs enfants est de qualité inférieure à celle des enfants d'autres pays, provinces ou districts. La classe politique et les décisionnaires répondent à la baisse qu'ils perçoivent dans la confiance du public en insistant encore plus sur l'emploi de tests standardisés comme mesure de la responsabilisation. Meaghan et Casas (1995b) disent que

> le testage est devenu la solution éclair, que l'administration des conseils scolaires et les représentants et représentantes des ministères de l'Éducation utilisent lorsqu'ils doivent produire des « résultats », et les produire rapidement. S'attaquer aux questions relatives au curriculum, améliorer la formation du personnel enseignant [et] élaborer des méthodes d'évaluation justes et efficaces ne donneraient

> pas de résultats assez rapidement, et les résultats obtenus ne seraient pas aussi simples à mesurer et à interpréter que les résultats des tests★ (p. 84).

Peut-on vraiment dire qu'il y a crise de confiance du public à l'égard du système scolaire? Les recherchistes Hart et Livingstone (1998), de l'IEPO/UT, ne sont pas de cet avis. C'est la question qu'ils se sont efforcés d'éclaircir dans leur analyse des tendances qui se dégagent de sondages d'opinion sur les attitudes vis-à-vis des écoles, réalisés en Ontario et dans tout le pays depuis 1980. Leur étude portait également sur la question de savoir si le public était « en faveur d'un programme d'austérité fiscale et de privatisation comme celui que prescrit la tendance néoconservatrice pour l'éducation »★ (p. 1). [9] Leurs recherches révèlent que les attitudes des chefs d'entreprises et du grand public sont nettement divergentes en ce qui concerne la confiance envers le système scolaire.

Plus précisément, Hart et Livingstone ont trouvé qu'il y avait peu de raisons de parler d'une « crise imminente dans la confiance du public ou des parents à l'égard des écoles »★ (p. 1). Cette conclusion contraste avec les « tendances que l'on retrouve chez les chefs d'entreprises, qui sont d'avis que la confiance envers le système scolaire est véritablement en crise et qui, en même temps, épousent les principes du programme néoconservateur » (p. 4), y compris la réduction des dépenses et le recours aux mécanismes du marché, programme pour l'éducation « qui ne semble avoir qu'un appui limité auprès du grand public »★, selon les auteurs (p. 16).

[9] Les auteurs définissent le néoconservatisme en général comme étant « une position politique de la Nouvelle Droite [...] qui préconise : a) la réduction du rôle de l'état, particulièrement en ce qui concerne les services sociaux, la redistribution des revenus et les initiatives en faveur de l'équité; b) la réduction de la dette et des déficits gouvernementaux; c) le recours aux mécanismes du marché et à la concurrence, par la privatisation et l'augmentation de la concurrence dans le domaine public, afin de rendre les services publics plus efficaces »★ (p. 2).

Les tests, les écoles et la concurrence du marché

Le secteur des entreprises est très en faveur de normes nationales en matière d'éducation et de tests nationaux. Grâce au lobbisme efficace exercé par des organismes représentant de puissants intérêts commerciaux, comme le Conseil canadien des chefs d'entreprises, l'Institut C.D. Howe et le Conference Board du Canada, les tests standardisés occupent une place privilégiée dans les programmes du gouvernement et du public. Les tests standardisés cadrent parfaitement avec les priorités des entreprises — la correction informatisée des tests à choix multiples coûte moins cher et est plus rapide que la correction de dissertations ou d'autres travaux écrits (Barlow et Robertson, p. 118); les résultats des tests standardisés correspondent, dans le domaine de l'éducation, aux compétences minimales essentielles exigées des entreprises (Reardon, Scott et Verre, 1994, p. 2); et comme nous le signalions plus tôt, le testage en éducation peut être source de profits considérables.

Les entreprises appuient aussi les tests standardisés en éducation parce que ces derniers sont conformes à une idéologie de marché libre. Accorder le choix d'écoles aux parents, considérés comme des consommateurs et consommatrices de l'éducation, est préconisé comme un moyen de promouvoir la concurrence entre les écoles. La concurrence est perçue comme un incitatif nécessaire pour assurer une éducation de qualité (les écoles « inefficaces » ferment leurs portes tandis que les « bonnes » écoles savent relever le « défi du marché »). Dans ce contexte, il faut avoir une base permettant de comparer les écoles et de prendre des décisions : c'est là qu'entrent en jeu le classement et la publication des résultats moyens obtenus aux tests standardisés (Berthelot, 1995, p. 4). Bien entendu, le classement des écoles (ou même des districts, des provinces ou des états) à partir de résultats de tests est une pratique qui laisse grandement à désirer. Non seulement les résultats sont souvent présentés hors contexte, mais une augmentation minime du résultat d'élèves particuliers peut modifier considérablement le classement

national selon un barème, et cette modification peut facilement être manipulée à des fins politiques. Si on ajoute à cela les fortes pressions exercées pour accroître les notes obtenues aux tests et les méthodes douteuses utilisées à cette fin, une manipulation accrue des résultats et la publication de renseignements trompeurs deviennent très probables.

Dans un système de libre choix des écoles (y compris des écoles à charte), les écoles dites « bonnes » attirent et sélectionnent les meilleurs élèves. Les écoles publiques locales commencent à ressentir l'effet négatif de cet « écrémage » par les écoles d'élite des élèves motivés et plus performants et de leurs parents plus activement engagés dans l'éducation, et se retrouvent avec moins d'élèves (ceux qui sont à risque élevé et qui coûtent le plus cher) et donc avec moins d'argent des contribuables (Dehli, 1998, p. 31).

Selon la Fédération canadienne des enseignantes et des enseignants (1997),

> la concurrence entre les écoles crée de fortes demandes pour les élèves qui, déjà, ont les meilleures chances de réussir dans nos écoles. Elle porte les écoles à éviter de servir les enfants en difficulté et les enfants handicapés qui peuvent nécessiter plus que leur « part » des ressources, sans toutefois rehausser le profil de l'école dans des domaines clés de la commercialisation comme les résultats aux tests standardisés (p. 17).

Il convient aussi de remarquer que lorsque les élèves plus performants sont attirés par des programmes plus exigeants et y sont concentrés, on crée peut-être ainsi l'illusion d'un meilleur rendement. Fuller et Elmore (cités dans Dehli, 1998, p. 31) nous mettent en garde : « la possibilité d'avoir le choix aura probablement pour seul résultat de déplacer d'une école à l'autre les élèves les meilleurs, donnant la fausse impression que la concentration d'élèves forts et motivés tient à la qualité de l'école ou au système de libre choix d'école. »

La « pancanadianisation » du curriculum et de l'évaluation

Le programme pancanadien (c.-à-d. national) d'action en éducation piloté par le Conseil des ministres de l'Éducation (Canada) (CMEC) et présenté dans la « Déclaration de Victoria » de 1993 prévoit d'importantes initiatives aux chapitres du curriculum, de l'évaluation des élèves et des indicateurs scolaires. En ce qui concerne le curriculum, le CMEC publiait en 1997 son *Cadre commun de résultats d'apprentissage en sciences de la nature : Protocole pancanadien pour la collaboration en matière de programmes scolaires*. Il s'agit d'un cadre pour l'élaboration d'un tronc commun qui établit les résultats d'apprentissage souhaités dans l'enseignement des sciences de la maternelle à la 12e année (pour toutes les provinces sauf le Québec). Robertson (1998b) fait remarquer que le « CMEC a escamoté l'épineuse question constitutionnelle de la responsabilité des provinces en matière de curriculum, en déclarant que le document n'était qu'un cadre destiné aux personnes chargées d'élaborer les curriculums, et qu'il ne s'agissait pas d'un curriculum à proprement parler. Mais, personne n'a été dupe. »* (p. 86). Elle constate de plus que la démarche de consultation a également été éludée. Les délais accordés pour formuler des commentaires sur les ébauches de ce long document étaient serrés et la nature des commentaires était très circonscrite — « les personnes qui souhaitaient faire des commentaires ne pouvaient mettre en question ni le bien-fondé d'un curriculum national ni son attachement contesté aux résultats »* (p. 86). Les organisations d'enseignantes et d'enseignants, les associations de conseils scolaires et les autres organismes qui ont toujours participé à l'élaboration des curriculums ont effectivement été exclus de la démarche.

À l'échelle régionale, des curriculums sont aussi en cours d'élaboration par l'entremise de la Fondation d'éducation des provinces Atlantiques, dans l'Est, et du Protocole de collaboration concernant l'éducation de base dans l'Ouest canadien. Ces initiatives de réforme des curriculums sont conformes à l'importance croissante

que l'on accorde aux résultats des tests standardisés et aux matières de base comme les mathématiques, les sciences et les compétences langagières (Dunning, 1997).

Maintenant que des programmes d'évaluation provinciaux sont en place dans la plupart des provinces et des territoires, un programme national complet d'évaluation ne devrait pas tarder. Barker (1991) analyse le Programme d'indicateurs du rendement scolaire (PIRS)[10], projet national entrepris en 1989 par le CMEC dans le but d'évaluer les élèves canadiens de 13 à 16 ans en mathématiques, en lecture et écriture, et en sciences. Barker (1991) soutient que ce projet est une initiative politique mise en œuvre pour répondre aux critiques formulées à l'égard de l'éducation et aux demandes faites par les entreprises, les industries et les gouvernements pour une responsabilisation accrue. Elle suggère que le PIRS peut être considéré comme un catalyseur de la réforme de l'éducation au Canada, pouvant aller de changements plus restreints au niveau de l'institutionnalisation des tests nationaux jusqu'à des transformations radicales comme un contrôle accru de l'éducation par le gouvernement fédéral et l'institutionnalisation de normes nationales en matière d'éducation dans un but de productivité économique. Le rapport de 1994 du Comité des normes nationales en éducation du Conseil consultatif national des sciences et de la technologie y fait écho. Le Comité a notamment recommandé au premier ministre :

[10] Pendant le premier cycle du PIRS, des évaluations ont été faites en mathématiques (1993), en lecture et en écriture (1994) et en sciences (1996). Le second cycle du PIRS a comporté des évaluations en mathématiques en 1997, en lecture et écriture en 1998 et en sciences en 1999. Le PIRS est financé par les provinces, par Développement des ressources humaines Canada au niveau fédéral et par des entreprises privées (Groupe Investors Inc. et Spar Aérospatiale Limitée). Le CMEC estime à environ 11 millions de dollars les coûts totaux d'administration et de mise en oeuvre du Programme, depuis l'étape de planification en 1989 à la réalisation complète en 1999, non compris les coûts de l'appui non financier des provinces [renseignements tirés du site Web du CMEC, http://www.cmec.ca/saip/]. Même s'il existe depuis une décennie, le PIRS et son incidence semblent avoir fait peu l'objet d'analyses critiques.

√ « Que l'administration fédérale aide à l'établissement et à l'administration de tests *pour toutes les matières de base* [italiques ajoutés] en s'inspirant des travaux du Programme d'indicateurs du rendement scolaire. »

√ « Que la fréquence de l'administration des tests concernant toutes les matières soit accrue de manière à maximiser l'information nécessaire pour assurer une amélioration continue. Si des tests annuels permettent d'obtenir un avantage maximum, il faudrait y avoir recours » (p. 25).

Le Comité a aussi demandé l'établissement de normes nationales en éducation et, dans une section portant sur le besoin de tests pour mesurer les compétences relatives à l'employabilité, recommande que « l'évaluation des compétences [d'employabilité] de base, une fois définies, figure dans le PIRS à mesure que son mandat sera établi et élargi » (p. 25).

Christensen (1998) a analysé le CMEC depuis ses débuts au sein de l'Association canadienne d'éducation jusqu'à son émergence en tant que, selon ses propres mots, « tribune nationale sur les politiques en éducation », et documente l'influence croissante d'un organisme qui, dit-elle, « a échappé en grande mesure à toute enquête critique depuis sa naissance il y a 30 ans » (p. iii), ne rend compte à aucun gouvernement et, malgré qu'il se soit officiellement engagé à consulter d'autres organismes éducationnels et à collaborer avec eux, « tend à exclure les autres par principe »★. Dans le chapitre qu'elle consacre aux travaux du CMEC relatifs au curriculum et à l'évaluation des élèves (les premières initiatives à ce sujet, dans les années 1980, ont marqué le début de l'influence croissante du CMEC), Christensen explique comment les pressions exercées par le gouvernement fédéral au début de la décennie ont « précipité la mise en œuvre du PIRS par le Conseil en 1993 » (p. 83). Le désir du gouvernement fédéral de remédier « à l'insatisfaction du public face à la "crise" de l'éducation au Canada » s'est exprimé lors d'un discours du trône prononcé en 1991, dans lequel le premier ministre

mentionnait le besoin de buts nationaux en matière d'éducation, marquant, comme le note Christensen, un retour au dossier fédéral sur les politiques "nationales" en matière d'éducation »* (p. 83). (Il est intéressant de signaler que presque à la même époque, le gouvernement américain publiait sa stratégie America 2000 en matière d'éducation, qui prévoyait un système de normes et d'évaluation nationales.) Attentif à ne pas empiéter sur les compétences provinciales, Ottawa a choisi de traiter des problèmes d'éducation par l'intermédiaire du Forum canadien du savoir, mis sur pied par le Comité directeur de la prospérité (Initiative de la prospérité), dans le but « d'entamer un processus visant à énoncer les attentes en matière d'éducation et de formation au Canada »* (p. 84). À l'occasion d'une réunion du « Forum du savoir » organisée par le Secrétariat de la prospérité et le Conference Board du Canada en 1993, le député fédéral Bernard Valcourt a déclaré que l'Initiative de la prospérité aura permis de « lier dans l'esprit des Canadiens et des Canadiennes la compétitivité et l'apprentissage »* (Robertson, citée dans Christensen, 1998, p. 86). Christensen semble croire que la critique du Comité directeur fédéral de la prospérité à l'égard des activités du CMEC est un des facteurs qui ont fait avancer le dossier du PIRS.

L'accent mis par le CMEC sur la responsabilisation par l'évaluation résulte en grande partie de la « quasi-obsession » actuelle à l'égard de la concurrence sur le marché mondial et de l'anxiété qu'elle suscite, ainsi que du besoin qui en résulte de savoir « exactement où l'on se situe par rapport à la concurrence, aux niveaux national et international, condition sine qua non de l'entrée en combat »* (Christensen, 1998, p. 76). Bien qu'il n'ait pas réagi aux préoccupations exprimées au sujet du PIRS par des organismes éducationnels, y compris des organisations d'enseignantes et d'enseignants, le CMEC a « choisi de privilégier certaines perspectives »* (p. 92), entre autres les intérêts des entreprises.

Dans un projet d'énoncé de position sur la responsabilisation et le testage en éducation, la Fédération canadienne des enseignantes et des enseignants (1992) soulevait plusieurs questions importantes

concernant les réactions possibles de diverses administrations aux premiers résultats du projet national de testage du CMEC et aux répercussions profondes du PIRS à plus long terme. Selon ce document :

> Il va de soi que certaines provinces se classeront moins bien que d'autres (en fait, on peut presque prédire les résultats, si l'on se fonde sur les données déjà existantes). Que feront ces provinces? Décideront-elles d'adopter le curriculum et les stratégies organisationnelles des provinces qui ont mieux réussi (ce qui contribuerait à promouvoir l'adoption d'un tronc commun national)? Essaieront-elles de modifier leur curriculum actuel pour qu'il corresponde davantage aux questions des tests? Les conseils scolaires chercheront-ils à répéter les tests, ce qui pourrait faciliter l'identification d'écoles et d'enseignantes et d'enseignants particuliers? Encouragera-t-on le personnel enseignant à adapter leur enseignement à la prochaine batterie de tests? Le programme de testage donnera-t-il naissance à un système scolaire médiocre et surstandardisé où la créativité et l'autonomie des élèves et du personnel enseignant n'ont plus leur place, et où les objectifs les plus négligés sont précisément les buts moins définis de l'éducation que tous s'accordent à considérer comme particulièrement importants?* (p. 44).

Le PIRS en est presque à la fin de son deuxième cycle complet et des préoccupations commencent à surgir au sujet de certains aspects du programme (la méthode servant à fixer les attentes concernant le rendement des élèves, par exemple). Comme Robertson (1999) le fait remarquer, il semble que le Programme influe dans une « mesure considérable sur l'homogénéisation des curriculums, dans un pays où l'autorité locale en matière de curriculum a toujours été considérée comme un bien plutôt qu'un mal »* (p. 715). La création de curriculums axés sur les résultats et conformes au PIRS par des

consortiums régionaux des provinces de l'Ouest et de l'Atlantique s'inscrit dans cette tendance.

Le document de la FCE poursuit en soulignant l'importance de disposer de moyens éducatifs adéquats pour réaliser des objectifs vastes et diversifiés en matière d'éducation :

> L'erreur fondamentale du projet du CMEC est qu'il ne reconnaît pas que les besoins en matière de testage doivent être dictés par les buts de l'éducation, le curriculum et l'enseignement. Autrement, les résultats des tests risquent de devenir une fin en soi et de donner lieu à des pratiques indésirables dans les écoles [...] Même si les tests contribuent à l'avancement et à l'amélioration de l'éducation, il semble que le Programme d'indicateurs du rendement scolaire du CMEC appartienne à la catégorie des tests conçus principalement comme outils de surveillance et de contrôle* (p. 44 et 46).

Selon Meaghan et Casas (1995b), le PIRS est un exemple typique de ce qui arrive lorsqu'on « met la charrue avant les bœufs » — un programme national de testage est mis au point alors qu'il n'existe pas de consensus public sur les buts éducationnels que nous devrions viser et sur la façon dont ils devraient être reflétés dans le curriculum. Il en découle qu'« une poignée de bureaucrates et d'experts-conseils »* décideront de ce que les élèves doivent savoir (p. 88).

Comme nous le disions auparavant, le curriculum est de plus en plus axé sur les tests et répond de moins en moins aux besoins des élèves et de la collectivité. Les moyens employés en éducation dictent des objectifs définis de plus en plus étroitement. Barlow et Robertson font remarquer que les tests standardisés

> contribuent fortement à promouvoir la réforme du curriculum et l'élaboration de buts éducationnels. Les éléments ayant fait l'objet de tests et ceux qui le feront à l'avenir deviennent prioritaires, non seulement dans

> l'esprit du public ou dans les politiques ministérielles, mais aussi en fait de temps et d'attention qu'on y consacre en classe. Les autres priorités passent au second plan et l'équilibre du curriculum change en douceur et sans avoir suscité de débat★ (p. 118 et 119).

C'est d'ailleurs ce qui se produit chez nos voisins du Sud. Au cours des 150 dernières années, les exigences en matière de testage aux États-Unis ont eu des conséquences frappantes, bien que non intentionnelles, sur les politiques éducationnelles. Les curriculums ont notamment été remaniés et les attitudes concernant les buts de l'éducation ont changé (Madaus, 1994, p. 79). L'influence sur l'éducation d'institutions comme l'Educational Testing Service (ETS), la plus importante entreprise privée de tests au monde, est considérable et ne peut être sous-estimée. Dans un article publié dans *Atlantic Monthly*, Lemann (1995) déclare que

> l'ETS et ses tests sont devenus une obsession nationale, profondément ancrée dans la vie de la classe moyenne. Il s'est créé une importante industrie indépendante de préparation aux tests qui a fortement influencé l'éducation élémentaire et secondaire, dont l'un des objectifs principaux est la préparation aux tests de l'ETS★ (p. 42).

Une vision d'ensemble de la réforme de l'éducation

Les tendances en matière d'évaluation énoncées dans l'introduction du présent rapport doivent être placées dans le contexte de tendances plus générales à l'égard des changements aux systèmes d'éducation, notamment : la réduction du financement, des réformes que l'on justifie de plus en plus pour des raisons économiques (c.-à-d., l'éducation est un moyen de préparer les élèves à réussir sur le marché du travail et les pays à affronter la concurrence mondiale) et un

penchant à adopter des solutions axées sur le marché, selon lesquelles l'éducation est un produit et les élèves et leurs parents en sont les consommateurs et consommatrices (Levin, 1998). Ces tendances, qui sont davantage motivées par des considérations fiscales, politiques et idéologiques que par un désir sincère d'accroître l'accès à une éducation de qualité, n'étonnent guère dans un monde où prévaut l'idéologie de la mondialisation.[11]

De plus en plus l'influence d'organismes internationaux comme l'Organisation de coopération et de développement économiques (OCDE) et l'Organisation de coopération économique Asie-Pacifique (APEC) se fait sentir sur les politiques en matière d'éducation. Dans son analyse du document de principe de l'APEC sur l'éducation, préparé par le gouvernement de la Corée du Sud, Kuehn (1997b) fait remarquer que « la tendance à la mondialisation de l'éducation » repose sur le principe que « la mondialisation est inévitable et que l'éducation [tout comme les autres secteurs de la société] doit se conformer à ses exigences ». Comme l'éducation dispensée dans cette économie mondiale vise principalement à préparer les travailleurs et les travailleuses au monde des affaires, « les entreprises devraient jouer un rôle central dans la détermination du contenu des études ». Ce contenu se voit alors limité à l'éthique, aux attitudes et aux compétences « qui conviennent à une vie de travail ». Dans ce contexte, les tests standardisés sont perçus comme un moyen commode de réglementer la production du « capital humain » nécessaire. Kuehn (1997a) illustre cette réalité en donnant l'exemple du Mexique, où un examen standardisé, l'*examen único*, a été créé pour aiguiller un plus grand nombre d'élèves vers des programmes professionnels et techniques où sont enseignées les

[11] « On Globalization », de Larry Kuehn, Fédération des enseignantes et enseignants de la Colombie-Britannique, résume certaines des principales caractéristiques de la mondialisation [ce document peut être consulté sur Internet : http://www.ctf-fce.ca/e/what/other/onglobe.htm].

compétences requises par l'économie mondiale, et pour détourner les élèves de programmes universitaires « visant principalement le développement social et culturel du Mexique »★ (p. 86). Voilà qui a soulevé un tollé : en août 1996, des manifestants et manifestantes ont bloqué des rues passantes de Mexico et ont forcé le marché boursier mexicain à fermer pendant une journée complète.

L'analyse que fait Robertson (1998a) des tests à grands enjeux dans le contexte de la mondialisation illustre bien à quel point le testage est en passe de devenir un levier important pour la réalisation d'un certain nombre de réformes scolaires d'envergure. La mondialisation nous rapproche inexorablement de ce qu'elle appelle un « McMonde » — un monde très inéquitable et divisé en tiers : où « un tiers est constitué des personnes qui réussiront assez bien après la restructuration économique internationale, tandis que les deux autres tiers ne réussiront pas »★; et où « tout est conçu en fonction des besoins du marché et de la recherche effrénée de profits »★. Elle constate que « le système scolaire public est victime d'un mouvement pour la création de "McÉcoles" qui répondront aux besoins du McMonde », et elle montre dans quelle mesure les tests à grands enjeux peuvent servir de

> moyen de faire passer un certain nombre de réformes au service de McÉcoles que l'on rejetterait en d'autres occasions comme étant inconstitutionnelles, inopportunes ou contraires à tout ce que nous connaissons sur les moyens d'encourager les bonnes écoles et les bons élèves. Les tests à grands enjeux sont liés si étroitement et en même temps si obscurément aux autres réformes des McÉcoles — le rétrécissement du curriculum, la déprofessionnalisation du personnel enseignant, la privatisation et l'enthousiasme obligatoire à l'égard de la technologie — que l'emploi de cette stratégie est brillant.★

En plus d'incarner les principales valeurs culturelles associées à la concurrence, à l'efficience, au rendement et aux résultats (les plus concrets et mesurables possible), des mécanismes comme les tests

standardisés à grands enjeux et les indicateurs de rendement sont un lien essentiel entre les deux tendances parallèles et apparemment contradictoires de la réforme scolaire, la décentralisation (État faible) et la centralisation (État fort). Alors que la décentralisation se traduit par des réformes comme la gestion locale, selon laquelle la prise de décisions est déléguée aux écoles, et le libre choix d'école, en vertu duquel le pouvoir est remis aux « consommateurs et consommatrices » de l'éducation, la centralisation se réalise au chapitre du financement, des curriculums, des normes et de l'évaluation, graduellement pris en main par les gouvernements. Comme nous le faisions remarquer plus tôt, l'évaluation de résultats fixés centralement et la publication des notes obtenues lors de cette évaluation permettent aux parents de comparer les écoles et de faire un choix — ce qui facilite la mesure de la *responsabilisation* du système scolaire *face au marché*. Dans sa critique du livre de Whitty, Power et Halpin, *Devolution and Choice in Education: The School, the State and the Market*, Apple (1998) traite de la notion « d'État évaluatif » que présentent les auteurs :

> Même s'il semble déléguer un certain pouvoir à des personnes et à des institutions autonomes qui se font de plus en plus concurrence sur le marché, l'État demeure fort dans les secteurs clés [...] C'est d'ailleurs ce qui caractérise le néolibéralisme à bien des égards. En effet, ce qui distingue essentiellement le libéralisme classique et sa confiance dans « l'initiative privée » au sein du marché des formes courantes de néolibéralisme est son attachement à un État régulateur. Le néolibéralisme exige la production constante de preuves confirmant « le respect des normes d'une entreprise privée » [...]. Dans un tel contexte, non seulement l'éducation devient un produit du marché comme le pain et les voitures, dans lequel les valeurs, les méthodes et les métaphores du monde des affaires dominent, mais ses résultats doivent pouvoir se réduire à des « indicateurs de rendement » standardisés (cité dans Apple, 1998, p. 25).

Dans un système axé sur le marché, les tests et les indicateurs de rendement servent effectivement à démontrer que les écoles et les systèmes scolaires fonctionnent comme « des entreprises privées » et satisfont ainsi aux exigences de la reddition de comptes face au marché — aux dépens de la véritable responsabilisation envers le public.

Observations générales

Passant du contexte mondial à celui de la salle de classe, nous nous devons de souligner que le personnel enseignant reconnaît l'importance d'évaluer l'apprentissage des élèves et de trouver les meilleurs moyens de le faire en salle de classe. Les enseignantes et les enseignants mesurent régulièrement les progrès des élèves par diverses méthodes, qu'ils conçoivent et administrent eux-mêmes — examens de fin de trimestre, interrogations, rédactions, projets, portfolios et observation en classe, pour n'en nommer que quelques-unes. Selon Robertson (1998b), les diverses évaluations formelles et informelles réalisées par le personnel enseignant sont des « indicateurs remarquablement efficaces de la réussite future de leurs élèves, à la fois dans d'autres cours et dans la vie en général »* (p. 65). Lorna Earl (1995), professeure de théorie et d'études en matière de politiques éducationnelles à l'IEPO/UT, croit que « les enseignantes et enseignants en salle de classe et les évaluations qu'ils effectuent quotidiennement peuvent, en fait, constituer la base des liens de responsabilité les plus importants » (p. 6). Du fait de leurs contacts réguliers et de leurs rapports étroits avec les élèves, les enseignantes et les enseignants sont les mieux placés pour procéder à des évaluations continues, tenir les élèves (et leurs parents) au courant de leur rendement et recommander des moyens qui favoriseront l'apprentissage des élèves. Citant l'ouvrage de R. Stiggins, *Student-Centered Classroom Assessment*, Earl (1995) fait remarquer que ce qui justifie l'importance de l'évaluation faite en classe pour l'amélioration de l'apprentissage est aussi la raison pour laquelle les évaluations à grande échelle ne donnent pas d'aussi bons résultats :

> « [...] tous les programmes de testage centralisés partent du principe que les décisions qui ont le plus d'effet sur la qualité de l'école sont prises au niveau de la gestion de l'école et de l'élaboration des politiques; autrement dit, quelque part à l'extérieur de la salle de classe. [...] Il est naïf et, de toute évidence, nuisible de placer l'origine

> de l'amélioration de l'école à ce niveau. L'amélioration de l'école ne jaillit pas des coulisses de la politique. Trois groupes de décisionnaires déterminent la qualité et, par conséquent, l'impact des écoles américaines. Par ordre d'importance, il y a les élèves comme tels, leurs enseignantes et leurs enseignants, et leurs parents. Les décisions de ces trois groupes sont tributaires de la même source d'information pour ce qui est du rendement des élèves, et cette source ne se trouve pas dans les tests standardisés menés au niveau national ou local ou administrés par les états. Leurs décisions reposent toutes sur l'évaluation du rendement de l'élève que l'enseignante ou l'enseignant effectue au jour le jour. Pour cette raison, on n'améliore pas l'école en augmentant le nombre de tests standardisés, mais plutôt en élaborant et en utilisant les meilleurs instruments de mesure que nous puissions produire pour les évaluations en salle de classe! » (cité dans Earl, 1995, p. 8).

En plus de noter que le personnel enseignant joue un rôle central dans le processus de mesure et d'évaluation, nous devons aussi reconnaître la valeur des objectifs multiples de l'éducation — objectifs qui dépassent la simple préparation des élèves à intégrer le marché du travail, aussi important que cela puisse être. Il faut également évaluer le degré de réalisation de *tous* ces objectifs, et non seulement les aptitudes cognitives, pour améliorer la qualité de l'éducation. Même si l'évaluation efficace d'objectifs comme la préparation des élèves à assumer leurs devoirs civiques est une tâche complexe et importante, « c'est précisément cette complexité qui fait que l'enseignement de ces objectifs est si important »* (Barlow et Robertson, p. 119).

Il est irresponsable de recourir à une seule forme d'évaluation, surtout une méthode aussi imparfaite que les tests standardisés, pour prendre des décisions importantes concernant les élèves, le personnel enseignant et les écoles. Il est tout aussi dangereux de se servir d'une

seule forme d'évaluation à des fins diverses (le progrès des élèves, le rendement des enseignantes et des enseignants ou l'efficacité des programmes et du système). Il faut utiliser les instruments qui conviennent à chacun des objectifs visés (Earl, 1995, p. 7).

Nous avons besoin de réexaminer le concept de la responsabilisation et sa signification véritable. Le testage en soi ne constitue pas une forme de responsabilisation, contrairement à ce que nous feraient croire les tendances courantes relatives aux politiques d'évaluation. Darling-Hammond (1999) rappelle que

> la responsabilisation est un moyen de faire en sorte que les individus ou les organisations répondent de leurs actions afin que les personnes qui en dépendent soient assurées de l'existence de mesures de protection visant à encourager de bonnes pratiques et à prévenir des pratiques indésirables ou abusives, à ménager des recours en cas de problèmes et à offrir une certaine garantie de traitement juste et équitable.

Dans cette optique, elle signale que dans un véritable système complet de responsabilisation en matière d'éducation, les divers intervenants, allant des gouvernements fédéral et provinciaux jusqu'à l'école et à l'enseignant ou l'enseignante, ont des rôles et des responsabilités d'importance à l'égard desquels ils doivent être tenus de rendre des comptes. Par exemple, l'état ou la province « devrait avoir la responsabilité de munir les écoles des ressources voulues pour fournir des services d'éducation satisfaisants, de veiller à ce que ces ressources soient réparties de façon équitable et que toutes les écoles dans ce contexte soient dotées d'un personnel qualifié et bien formé. »* Il appartient aux districts scolaires locaux d'assurer « des pratiques qui contribuent à la qualité de l'enseignement et de l'apprentissage dans les écoles, et des procédés qui appuient de telles pratiques. »* Les enseignantes et les enseignants ont la responsabilité de « prendre de bonnes décisions éducationnelles au nom des enfants. »* Darling-Hammond (1999) décrit cette dernière comme

la responsabilité professionnelle, qui suppose un investissement dans le personnel enseignant, le savoir et le savoir-faire. Voilà qui est une composante essentielle d'un système de responsabilisation véritable, qui exige que l'on donne accès aux enseignantes et aux enseignants de façon continue aux connaissances et à l'appui nécessaires aux prises de décisions judicieuses afin d'améliorer l'apprentissage chez les élèves. À moins que *tous* les intervenants soient tenus de rendre compte de leurs diverses responsabilités, ils ne peuvent parler de responsabilisation que dans le vide.

En résumé, du point de vue de l'apprentissage, peu de preuves démontrent l'utilité des tests standardisés. En fait, ces tests nuisent à l'apprentissage, particulièrement dans le cas des élèves à risque — car ils perpétuent et accentuent les inégalités dans le système scolaire. L'évaluation, y compris l'évaluation authentique, doit avant tout s'inscrire dans une stratégie diversifiée visant à améliorer la qualité de l'éducation et à corriger les iniquités par le biais d'autres réformes systémiques qui tiennent compte des conditions sociales et scolaires dans leur ensemble. Si notre système scolaire a pour principal objectif d'offrir des chances égales à tous les élèves, les tests standardisés ne serviront qu'à nous éloigner du but.

Annexe A — Sommaire des problèmes que posent les tests standardisés

D'abondantes recherches ont fait ressortir les nombreux problèmes associés à la nature et à l'emploi des tests de rendement standardisés. La liste suivante n'est que partielle (tirée de Meaghan et Casas, « On the Testing of Standards and Standardized Achievement Testing: Panacea, Placebo, or Pandora's Box? », *Interchange*, vol. 26, n° 1, 1995, p. 34 et 35) :

√ Afin de permettre la correction automatisée, les tests standardisés doivent se limiter à des questions à choix multiples, ce qui restreint considérablement la matière qui peut être évaluée. Par exemple, on peut évaluer les compétences langagières, mais non l'aptitude à rédiger ni la façon dont les élèves utilisent la langue.

√ Pour que les résultats demeurent comparables sur une longue période, on n'apporte que très peu de modifications substantielles d'une année à l'autre, ce qui encourage peu la mise à jour du curriculum.

√ Même si ces tests sont préconisés comme moyen d'établir des comparaisons internationales sur l'apprentissage et le rendement des élèves, ils ne tiennent pas compte des différences dans les curriculums selon l'âge et l'année d'études dans les divers pays (et ne peuvent en tenir compte).

√ Comme la même valeur numérique est accordée à chaque question, l'élève qui répond correctement à un plus grand nombre de questions difficiles (ou importantes du point de

vue du curriculum) est considéré l'égal de celui qui ne répond correctement qu'aux questions faciles.

√ Afin de disperser adéquatement les résultats [pour faciliter les comparaisons], on tend à éviter les questions auxquelles les élèves répondent presque toujours correctement ou incorrectement. Ainsi, certains aspects du rendement qui devraient être évalués risquent de ne pas être représentés du tout parce qu'ils ne contribuent pas à la dispersion des résultats. Les menus détails sont donc trop ciblés dans les tests et les notions importantes le sont insuffisamment.

√ Les tests standardisés mettent l'accent sur le produit de l'apprentissage et non sur le processus. Ils mesurent la capacité des élèves à mémoriser des faits, à définir des mots et à effectuer des calculs simples, et non des processus d'apprentissage de niveau supérieur tels que l'analyse, la synthèse, la formulation d'hypothèses et l'exploration de nouveaux moyens de résoudre des problèmes.

√ Fait plus crucial encore, les tests standardisés n'ont aucun rapport avec des normes de rendement : ils ne mesurent qu'en fonction de normes relatives.

√ Les tests standardisés ne peuvent servir d'outils diagnostiques utiles car ils ne montrent pas aux élèves où ils ont fait fausse route; les tests corrigés ne sont jamais remis aux élèves, ce qui empêche la rétroaction et la discussion qui sont au cœur de l'apprentissage et de l'éducation.

√ Les tests standardisés excluent systématiquement la créativité et l'esprit critique des élèves en éliminant le besoin de structurer leurs propres réponses.

√ Au lieu d'augmenter la responsabilisation, les tests standardisés ne font que la transférer du personnel enseignant et des autorités scolaires à des fonctionnaires anonymes ou à des bureaucrates de sociétés qu'on ne peut ni confronter ni tenir responsables de tests mal conçus, mal administrés ou mal corrigés.

√ De nombreuses études montrent que les tests standardisés sont entachés de préjugés contre des groupes minoritaires socioéconomiques, raciaux et ethniques.★

Annexe B —
Principes d'équité relatifs aux pratiques d'évaluation des apprentissages scolaires au Canada

(Comité consultatif mixte)

En 1993, un groupe de travail canadien formé par le Comité consultatif mixte du Centre for Research in Applied Measurement and Evaluation (University of Alberta) a publié un document intitulé *Principes d'équité relatifs aux pratiques d'évaluation des apprentissages scolaires au Canada*[12]. Ces principes et les lignes directrices connexes portent sur les évaluations faites en salle de classe et les évaluations standardisées à grande échelle préparées à l'extérieur de la classe (par des maisons d'édition spécialisées, des ministères de l'Éducation et des conseils scolaires), et représentent un large consensus du milieu de l'éducation (les organismes participants comprenaient entre autres l'Association canadienne d'éducation, l'Association canadienne des commissions/conseils scolaires et la Fédération canadienne des enseignantes et des enseignants). Destinés aux personnes qui conçoivent et utilisent les évaluations (celles-là mettent au point les méthodes et décident des politiques régissant les divers programmes

[12] Aux États-Unis, un projet du National Forum on Assessment, coalition formée d'organismes éducationnels et des droits de la personne, a donné lieu à la création en 1995 des *Principles and Indicators for Student Assessment Systems*. Le premier principe stipule que le but principal de l'évaluation est d'améliorer l'apprentissage des élèves.

d'évaluation; celles-ci choisissent et utilisent les méthodes d'évaluation, retiennent les services de concepteurs ou conceptrices ou prennent des décisions fondées sur les résultats et les conclusions des évaluations [p. 3]), les principes renferment notamment les points suivants :

√ Les méthodes d'évaluation devraient être adaptées aux buts de l'évaluation et à son contexte général.

√ Les élèves devraient avoir suffisamment d'occasions de manifester les connaissances, les compétences, les attitudes et les comportements qui font l'objet de l'évaluation.

√ Les procédés utilisés pour juger ou noter le rendement des élèves devraient être adaptés aux méthodes d'évaluation, et devraient être appliqués et contrôlés de façon systématique.

√ Les procédés employés pour faire la synthèse et l'interprétation des résultats devraient mener à des représentations précises et instructives du rendement d'un élève en rapport avec les buts et les objectifs d'apprentissage pour la période visée.

√ Les rapports d'évaluation devraient être clairs, précis et avoir une valeur pratique pour les personnes à qui ils s'adressent.

Annexe C — Principes de la FCE sur le curriculum, la mesure et l'évaluation

La Fédération canadienne des enseignantes et des enseignants a mis au point son propre ensemble de principes sur la mesure et l'évaluation. Au début de 1998, la FCE a organisé un atelier sur le curriculum et l'évaluation (dont il est question dans l'introduction) à l'intention de la direction des organisations d'enseignantes et d'enseignants à l'échelle du pays. Un des résultats importants de l'atelier a été la rédaction de projets de déclarations de principe sur la mesure et l'évaluation ainsi que sur le curriculum. Ces déclarations visent à orienter le travail des organisations provinciales et territoriales de l'enseignement dans un certain nombre de domaines, dont l'élaboration, l'application et la portée du curriculum, ainsi que les tests à grands enjeux et leur utilisation. Les déclarations de principe ont été adoptées en tant que principes directeurs de la FCE et sont reproduites ci-dessous :

Principes sur le curriculum

√ Le curriculum doit être polyvalent et holistique, afin de répondre aux besoins des enfants, des communautés et de la société en général.

√ Le curriculum doit refléter des attentes réalistes qui reposent sur la connaissance du développement de l'enfant.

- √ Le curriculum doit promouvoir, appuyer, accroître et illustrer l'équité et le respect à l'égard de la diversité.

- √ Le curriculum doit être conçu de manière à préparer les élèves à devenir des membres bienveillants, responsables et actifs de la société.

- √ Le curriculum est continu et dynamique.

- √ L'élaboration, l'application et l'évaluation du curriculum comportent de multiples étapes et, par conséquent, requièrent un engagement considérable de temps et de ressources.

- √ Les membres de la profession enseignante doivent être les principaux intervenants dans l'élaboration du curriculum.

- √ Le dialogue professionnel et l'action concertée sont essentiels à l'épanouissement et au perfectionnement professionnels.

- √ Les membres de la profession enseignante ont besoin de temps et de ressources pour s'adonner au perfectionnement professionnel et mettre en œuvre efficacement les modifications au curriculum.

- √ Les membres de la profession enseignante doivent avoir suffisamment d'autonomie pour pouvoir déterminer les méthodes, la fréquence des activités d'apprentissage et les modifications à apporter au curriculum, afin d'adapter ce dernier aux besoins des élèves.

- √ Les objectifs que fixe la société envers les élèves et les écoles doivent être stimulants, mais réalisables; le progrès accompli au regard de ces objectifs doit être mesuré de façon équitable et en fonction du contexte.

- √ Les conditions entourant l'enseignement et l'apprentissage se répercutent directement sur ce que les membres de la profession enseignante et les élèves peuvent accomplir ensemble.

- √ Le processus d'élaboration du curriculum doit donner l'occasion à la plus vaste représentation possible de la communauté d'exprimer son point de vue.

Principes sur la mesure et l'évaluation

- √ La mesure est la collecte d'informations fiables concernant la compréhension et les connaissances acquises par les élèves, y compris les ressources et les facteurs sociaux, économiques et éducationnels qui influent sur le rendement des élèves et du système.

- √ L'évaluation désigne la formulation de jugements fondés en partie sur les données issues de la mesure.

- √ La mesure et l'évaluation sont des démarches continues sur lesquelles s'appuient les décisions aux niveaux tant de la salle de classe que du système.

- √ Les enseignantes et les enseignants ont la responsabilité première en matière de mesure et d'évaluation du rendement des élèves.

- √ Les gouvernements et les systèmes scolaires ont une responsabilité collective de fournir aux enseignantes et aux enseignants des occasions de formation en poste en matière de mesure et d'évaluation.

- √ Les organisations de l'enseignement, en leur qualité de porte-parole de la profession, doivent participer directement à l'élaboration de toute politique sur la mesure et l'évaluation aux échelons provincial ou territorial, national et international.

- √ Le but de la mesure du rendement des élèves est de favoriser leur apprentissage.

- √ La conception, l'interprétation et l'application des informations qui dérivent de la mesure et de l'évaluation sont des tâches complexes liées aux convictions, aux attitudes et aux connaissances professionnelles.

- √ Les décisions par rapport à la mesure et à l'évaluation devraient porter sur ce qui est valable et ne pas se limiter à ce qui est mesurable.

- √ Les buts et les méthodes de mesure et d'évaluation de l'apprentissage des élèves, du curriculum, des programmes et des systèmes sont distincts et ne sont pas interchangeables.

- √ Les outils et les procédés utilisés pour mesurer et évaluer l'apprentissage des élèves doivent cadrer avec la théorie de l'apprentissage que reflète le curriculum.

- √ Les élèves devraient s'autoévaluer et évaluer leurs pairs, et être pleinement renseignés sur le but des activités de mesure et de l'évaluation.

- √ La mesure et l'évaluation des élèves, qu'elles soient réalisées dans la salle de classe ou à un autre niveau, doivent tenir compte des différences entre les sexes ainsi que de la diversité culturelle, linguistique et socioéconomique, entre autres.

- √ Il existe de nombreuses formes, méthodes et approches relatives à la mesure et aucune d'entre elles prise isolément ne suffit pour mesurer le rendement des élèves, des programmes ou des systèmes.

- √ Les renseignements provenant de la mesure et de l'évaluation du rendement des élèves devraient être recueillis, classés, communiqués et utilisés d'une manière qui respecte le droit des élèves à la protection de la vie privée.

Références

ALBERTA EDUCATION. « Details of $600 Million Boost for K-12 Education Announced », communiqué, Edmonton, http://ednet.edc.gov.ab.ca/news/1999nr/Mar99/nr-budgetpr.htm, 11 mars 1999.

« Alberta Proposes Bonuses for Student Achievement Gains. » *Canadian Principal*, 10(7), avril 1999, p. 1-4.

AMERICAN ASSOCIATION OF UNIVERSITY WOMEN EDUCATIONAL FOUNDATION, édit. *Gender Gaps: Where Schools Still Fail Our Children*, Washington D.C., 1998.

APPLE, M. « Are Markets and Standards Democratic? », critique de livre, *Educational Researcher*, vol. 27, n° 6, août/septembre 1998, p. 24-28.

BARKER, K. C. « Accountability as an Educational Reform Strategy: The Case of the School Achievement Indicators Program », document inédit, Département d'administration de l'éducation, University of Alberta, Edmonton, 10 décembre 1991.

BARLOW, M. et H.-j. Robertson. *Class Warfare: The Assault on Canada's Schools*, Toronto (Ontario), Key Porter Books, 1994.

BERTHELOT, J. *Gouverne démocratique ou dérives marchandes?*, communication préparée pour la Conférence nationale, « *L'éducation publique : Questions de l'heure* », de la Fédération canadienne des enseignantes et des enseignants, Montréal (Québec), mai 1995.

CASAS, F. et D. Meaghan. « Standardized Tests: Why They Cannot Measure Educational Quality », *FWTAO Newsletter*, vol. 13, n° 4, mars/avril 1995, p. 16-21.

CHRISTENSEN, L. « Governing Under an Exclusive Lock and Key: The Council of Ministers of Education, Canada, 1967-1997 », document inédit, Faculté des études supérieures et de la recherche (études canadiennes), Carleton University, Ottawa, 3 février 1998.

COMITÉ CONSULTATIF MIXTE, CENTRE FOR RESEARCH IN APPLIED MEASUREMENT AND EVALUATION. *Principes d'équité relatifs aux pratiques d'évaluation des apprentissages scolaires au Canada*, University of Alberta, Edmonton (Alberta), 1993.

CONSEIL CONSULTATIF NATIONAL DES SCIENCES ET DE LA TECHNOLOGIE. *Les normes nationales en matière d'éducation : Une question d'excellence. Rapport du Comité sur les normes nationales en éducation du Conseil consultatif national des sciences et de la technologie*, Ottawa, mai 1994.

DARLING-HAMMOND, L. (mars 1999). « Developing a Professional Model of Accountability for Our Schools. » [sur Internet] Discours prononcé à l'AGA de la Fédération des enseignantes et des enseignants de la Colombie-Britannique, mars 1999, http://www.bctf.bc.ca/parents/speeches/darling-hammond.html

DARLING-HAMMOND, L. « Performance-Based Assessment and Educational Equity », *Harvard Educational Review*, vol. 64, n° 1, printemps 1994, p. 5-30.

DARLING-HAMMOND, L. « The Implications of Testing Policy for Quality and Equality », *Phi Delta Kappan*, vol. 73, n° 3, novembre 1991, p. 220-225.

DEHLI, K. « Shopping for Schools: The Future of Education in Ontario? », *Orbit*, vol. 29, n° 1, 1998, p. 29-33.

DUNNING, P. *L'éducation au Canada : Vue d'ensemble*, Toronto (Ontario), Association canadienne d'éducation, juillet 1997.

EARL, L. M. *La responsabilisation en éducation : Où se situe le personnel enseignant?*, communication préparée pour la Conférence nationale, « *L'éducation publique : Questions de l'heure* », de la Fédération canadienne des enseignantes et des enseignants, Montréal (Québec), mai 1995.

EDUCATION WEEK ON THE WEB (en collaboration avec Pew Charitable Trusts). *Quality Counts '99: Rewarding Results, Punishing Failure*, Editorial Projects in Education [sur Internet] http://www.edweek.org/sreports/, 11 janvier 1999.

EISNER, E. « The Role of Teachers in Assessment and Education Reform » [sur Internet], discours prononcé à l'Assemblée générale annuelle de la Fédération des enseignantes et enseignants de la Colombie-Britannique, http://www.bctf.bc.ca/parents/speeches/eisner.html, mars 1998.

FAIRTEST (National Center for Fair & Open Testing), édit. *Selected Annotated Bibliography on the SAT (Scholastic Aptitude Test): Bias and Misuse*, Cambridge (Massachussetts), 1995.

FÉDÉRATION CANADIENNE DES ENSEIGNANTES ET DES ENSEIGNANTS, édit. *Derrière les mythes de l'école à charte*, document de recherche élaboré dans le cadre de la campagne des Dossiers nationaux en éducation, Ottawa (Ontario), septembre 1997.

FÉDÉRATION CANADIENNE DES ENSEIGNANTES ET DES ENSEIGNANTS. « Le rendement des écoles canadiennes : Les comparaisons internationales ont-elles leur utilité? » (fiche-débat n° 9 — Dossiers nationaux en éducation), octobre 1995a.

FÉDÉRATION CANADIENNE DES ENSEIGNANTES ET DES ENSEIGNANTS. « Des tests nationaux » (fiche-débat n° 1 — Dossiers nationaux en éducation), octobre 1995b.

FÉDÉRATION CANADIENNE DES ENSEIGNANTES ET DES ENSEIGNANTS. « Accountability and Testing in Education, Draft Position Paper », février 1992.

GARDNER, H. « Who Owns Intelligence? », *Atlantic Monthly*, vol. 283, n° 2, février 1999, p. 67-76.

GILLISS, G. *Programmes d'évaluation à l'échelle provinciale*, Ottawa (Ontario), Fédération canadienne des enseignantes et des enseignants, 1994.

GILLISS, G. « Why National Tests Won't Improve Education », *Globe and Mail*, 12 janvier 1993.

HALADYNA, T. M., S.B. Nolen et N.S. Haas. « Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution », *Educational Researcher*, vol. 20, n° 5, juin/juillet 1991, p. 2-7.

HART, D. et D.W. Livingstone. « The 'Crisis' of Confidence in Schools and the Neoconservative Agenda: Diverging Opinions of Corporate Executives and the General Public », *The Alberta Journal of Educational Research*, vol. XLIV, n° 1, printemps 1998, p. 1-19.

« Improvement, Not Incentives : Joint Proposal to Replace School Performance Incentive Program with Alberta School Improvement Program. » [sur Internet] Mémoire présenté collectivement au ministre de l'Éducation par la Alberta Home and School Councils' Association, The Alberta School Boards Association, l'Alberta Teachers' Association et le College of Alberta School Superintendents, mai 1999, http://www.teachers.ab.ca/what/asip.html

IRELAND, D. S. *Analysis of the Tasks Used in the Grade 3 Provincial Assessment of Mathematics and Language Arts 1997*, préparé pour l'Ontario English Catholic Teachers Association, 4 juin 1997a.

IRELAND, D. S. « Canadian Students Do Better Than Results Indicate », *CCPA Education Monitor*, vol. 1, nº 2, printemps 1997b, p. 13-16.

IRELAND, D. S. *The International Status of Canadian Education*, rév., Conseil d'éducation de Carleton, mai 1995.

JONES, K. et B.L. Whitford. « Kentucky's Conflicting Reform Principles: High-Stakes School Accountability and Student Performance Assessment », *Phi Delta Kappan*, vol. 70, nº 4, décembre 1997, p. 276-281.

KLEINSASSER, A. « Book Review — *Testing Testing: Social Consequences of the Examined Life* by F. Allan Hanson », *Harvard Educational Review*, vol. 64, nº 1, printemps 1994, p. 96-102.

KUEHN, L. « Parents, Standardized Testing and the Mexican Stock Exchange » *Our Schools/Our Selves*, vol. 8, nº 3, nº 51, avril/mai 1997a, p. 83-90.

KUEHN, L. « Schools for Globalized Business: The APEC Agenda for Education — A Commentary on the 'Concept Paper' for the APEC Human Resources Ministerial Meeting », [sur Internet], document de travail préparé pour le Forum sur l'éducation publique et la recherche, Sommet populaire de 1997 sur l'APEC, Vancouver (Colombie-Britannique), 19 et 20 novembre 1997, http://www.vcn.bc.ca/ed8resea/larry.html, 1997b.

LEMANN, N. « The Structure of Success in America », *Atlantic Monthly*, vol. 276, nº 2, août 1995, p. 41-60.

LEVIN, B. « An Epidemic of Education Policy: (What) Can We Learn From Each Other? », *Comparative Education*, vol. 34, nº 2, juin 1998, p. 131-141.

MADAUS, G. F. « A Technological and Historical Consideration of Equity Issues Associated with Proposals to Change the Nation's Testing Policy », *Harvard Educational Review*, vol 64, nº 1, printemps 1994, p. 76-95.

MADAUS, G. F. « The Effects of Important Tests on Students: Implications for a National Examination System », *Phi Delta Kappan*, vol. 73, nº 3, novembre 1991, p. 226-231.

MADAUS, G.F., & Kellaghan, T. (1993, April). "Testing as a Mechanism of Public Policy: A Brief History and Description." *Measurement and Evaluation in Counseling and Development*, vol. 26, nº 1, avril 1993, p. 6-10.

McLAUGHLIN, M. W. « Test-Based Accountability As a Reform Strategy », *Phi Delta Kappan*, vol. 73, n° 3, novembre 1991, p. 248-251.

McMAHEN, L. « Alberta to Grant Bonus Money to Boards for Performance », *The Edvisor*, vol. 1, n° 23, 15 mars 1999.

MCMURTRY, J. « National Testing », *Canadian Social Studies*, vol. 26, n° 3, printemps 1992, p. 94-95.

MEAGHAN, D. E. et F.R. Casas. « Let's Look Before We Leap: Standardized Testing and Special Education Students », *CEA Newsletter*, octobre/novembre 1995a, p. 5.

MEAGHAN, D. E. et F.R. Casas. « On Standardized Achievement Testing: Response to Freedman and Wilson and a Last Word », *Interchange*, vol. 26, n° 1, 1995b, p. 81-96.

MEAGHAN, D. E. et F.R. Casas. « On the Testing of Standards and Standardized Achievement Testing: Panacea, Placebo, or Pandora's Box? », *Interchange*, vol. 26, n° 1, 1995c, p. 33-58.

MEAGHAN, D. E. et F.R. Casas. « Quality Education and Other Myths: A New Face for an Old Conservative Agenda », *Our Schools/Our Selves*, vol. 7, n° 1, n° 43, septembre 1995d, p. 37-53.

MEDINA, N. et D.M. Neill. *Fallout From the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America's Public Schools*, 3e éd. rév., Cambridge (Massachussetts), National Center for Fair & Open Testing (FairTest), mars 1990.

NAGY, P. *Études comparatives internationales du rendement des élèves en mathématiques et en sciences*, Institut d'études pédagogiques de l'Ontario, février 1996.

NAGY, P. « International Comparisons of Student Achievement: Examining the Tests and the Data » (ébauche), communication, congrès annuel de la Société canadienne pour l'étude de l'éducation, Montréal, juin 1995.

NATIONAL FORUM ON ASSESSMENT. *Principles and Indicators for Student Assessment Systems*, Cambridge (Massachussetts), National Center for Fair & Open Testing (FairTest), 1995.

NEILL, M. et autres. *Testing Our Children: A Report Card on State Assessment Systems*, Cambridge (Massachussetts), National Center for Fair & Open Testing (FairTest), septembre 1997.

OAKES, J. *Keeping Track: How Schools Structure Inequality*, New Haven (Connecticut), Yale University Press, 1985.

PEIRCE, B. N. et P. Stein. « Why the 'Monkeys Passage' Bombed: Tests, Genres, and Teaching », *Harvard Educational Review*, vol. 65, n° 1, printemps 1995, p. 50-65.

REARDON, S. F., K. Scott et J. Verre. « Symposium: Equity in Educational Assessment — Introduction », *Harvard Educational Review*, vol. 64, n° 1, printemps 1994, p. 1-4.

REHNBY, N. *Une crise inventée : Effritement de la confiance dans l'éducation publique au Canada*, préparé par la Fédération des enseignantes et enseignants de la Colombie-Britannique à l'intention de la Fédération canadienne des enseignantes et des enseignants, février 1996.

ROBERTSON, H.-j. (mai 1999) « Bogus Points » [chronique « In Canada »], *Phi Delta Kappan*, 80(9), p. 715-716.

ROBERTSON, H.-j. « McSchool Meets McWorld: The Role of High-Stakes Testing », document préparé pour la Manitoba Teachers' Society, 5 mars 1998.

ROBERTSON, H.-j. *No More Teachers, No More Books: The Commercialization of Canada's Schools*, Toronto (Ontario), McClelland & Stewart, 1998b.

ROBERTSON, H.-j. *Le Cahier d'IDÉES amélioré : Liens entre la catégorie de sexe, la culture, les sciences et les écoles*, Ottawa (Ontario), Fédération canadienne des enseignantes et des enseignants, novembre 1992.

ROSSER, P. et autres. *Sex Bias in College Admissions Tests: Why Women Lose Out* (4e édition), Cambridge (Massachussetts), National Center for Fair & Open Testing (FairTest), août 1992.

SCHWARTZ, T. « The Test Under Stress », *New York Times Magazine*, 10 janvier 1999.

THURMOND, M. « The FairTest Case: Nothing Special, Just Equality », *Civil Liberties* (bulletin national de l'American Civil Liberties Union), n° 380, printemps 1994.

TORRANCE, H. « Combining Measurement-Driven Instruction with Authentic Assessment: Some Initial Observations of National Assessment in England and Wales », *Educational Evaluation and Policy Analysis*, vol. 15, n° 1, printemps 1993, p. 81-90.

152

TRAUB, R. *Les tests normalisés au Canada : Enquête sur les tests normalisés de rendement scolaire administrés par les ministères de l'Éducation et les commissions scolaires*, Toronto (Ontario), Association canadienne d'éducation, 1994.

WEISS, J. G. « It's Time to Examine the Examiners », *Testing African American Students — Special Re-Issue of The Negro Educational Review*, A. G. Hilliard III, Morristown (New Jersey), Aaron Press, 1991, p. 107-124.

WIDEEN, M. F., T. O'Shea, I. Pye et G. Ivany, « High-Stakes Testing and the Teaching of Science », *Canadian Journal of Education*, vol. 22, n° 4, 1997, p. 428-444.